INTELIGENCIA ARTIFICIAL: JUGAR O ROMPER LA BARAJA

Margarita Padilla García

útiles
traficantes de sueños

Desde sus inicios **Traficantes de Sueños** ha apostado por licencias de publicación que permiten compartir, como las Creative Commons, por eso sus libros se pueden copiar, distribuir, comunicar públicamente y descargar desde su web. Entendemos que el conocimiento y las expresiones artísticas se producen a partir de elementos previos y contemporáneos, gracias a las redes difusas en las que participamos. Están hechas de retazos, de mezclas, de experiencias colectivas; cada persona las recompone de una forma original, pero no se puede atribuir su propiedad total y excluir a otros de su uso o replicación.

Sin embargo, «cultura libre» no es sinónimo de «cultura gratis». Producir un libro conlleva costes de derechos de autor, traducción, edición, corrección, maquetación, diseño e impresión. Tú puedes colaborar haciendo una donación al proyecto editorial; con ello estarás contribuyendo a la liberación de contenidos.

Puedes hacer una donación
(si estás fuera de España a través de PayPal),
suscribirte a la editorial
o escribirnos un mail

Inteligencia artificial: jugar o romper la baraja

traficantes de sueños

Traficantes de Sueños no es una casa editorial, ni siquiera una editorial independiente que contempla la publicación de una colección variable de textos críticos. Es, por el contrario, un proyecto, en el sentido estricto de «apuesta», que se dirige a cartografiar las líneas constituyentes de otras formas de vida. La construcción teórica y práctica de la caja de herramientas que, con palabras propias, puede componer el ciclo de luchas de las próximas décadas.

Sin complacencias con la arcaica sacralidad del libro, sin concesiones con el narcisismo literario, sin lealtad alguna a los usurpadores del saber, TdS adopta sin ambages la libertad de acceso al conocimiento. Queda, por tanto, permitida y abierta la reproducción total o parcial de los textos publicados, en cualquier formato imaginable, salvo por explícita voluntad del autor o de la autora y solo en el caso de las ediciones con ánimo de lucro.

Omnia sunt communia!

útiles 33

Útiles es un tren en marcha que anima la discusión en el seno de los movimientos sociales. Alienta la creación de nuevos terrenos de conflicto en el trabajo precario y en el trabajo de los migrantes, estimula la autorreflexión de los grupos feministas, de las asociaciones locales y de los proyectos de comunicación social, incita a la apertura de nuevos campos de batalla en una frontera digital todavía abierta.

Útiles recoge materiales de encuesta y de investigación. Se propone como un proyecto editorial autoproducido por los movimientos sociales. Trata de poner a disposición del «común» saberes y conocimientos generados en el centro de las dinámicas de explotación y dominio y desde las prácticas de autoorganización. Conocimientos que quieren ser las herramientas de futuras prácticas de libertad.

1ª edición: 1250 ejemplares, octubre de 2025

Título: Inteligencia artificial: jugar o romper la baraja

Autora: Margarita Padilla García

Corrección ortotipográfica: Jabuti

Maquetación y diseño de cubierta:
Traficantes de Sueños
taller@traficantes.net
Edición:
Traficantes de Sueños
C/ Duque de Alba 13
28012 Madrid. Tlf: 915320928
e-mail:editorial@traficantes.net

@traficantes-ed.bsky.social
@traficantes_libreria
@traficantes_Ed

@TRAFICANTES_ED

ISBN: 978-84-19833-46-4
Depósito legal: M-18933-25

Inteligencia artificial: jugar o romper la baraja

Margarita Padilla García

traficantes de sueños
útiles

A los descubrimientos que revelan nuevas ignorancias

La máquina, obra de organización, de información, es, como la vida y con la vida, lo que se opone al desorden […]. La máquina es aquello por medio de lo cual el hombre se opone a la muerte del universo; hace más lenta, como la vida, la degradación de la energía, y se convierte en estabilizadora del mundo.

Gilbert Simondon,
El modo de existencia de los objetos técnicos

Agradecimientos

Agradezco la generosidad de Adriana, Amador, Carles, Carolina, Charo, Cristina, Emanuele, Jabuti, Lidia, Mar, Marta, Pablo, Raquel, Santi, otro Santi, Susana, Toni y Xabier por sus aportaciones para mejorar la calidad de este texto.

Índice

Presentación

Esto no es un libro de tesis. No pretende decir a nadie lo que tiene que pensar o lo que tiene que hacer, pues ni yo misma lo sé. Una persona puede tener opiniones, pero las tesis, ya lo sabemos, han de ser apuestas colectivas.

Mi intención al escribirlo ha sido abrir foco, aumentar, añadir complejidad, aportar dimensiones…

Lo he concebido como una baraja de naipes. Creo que en el «juego» de la inteligencia artificial, para llevarse alguna baza, hay que tener buenas cartas. Así que me he preguntado cuáles serían las que darían juego en las partidas a favor de la justicia y la equidad, y también a favor de la creatividad y la experimentación.

En las barajas los naipes van sueltos, pero al mismo tiempo tienen estructura. En este libro, cada capítulo es como un naipe. He intentado mantener un cierto orden, pero sé que habría sido posible barajarlos de muchas otras formas.

No todo el mundo se manejará igual de bien con todos ellos. Habrá quien prefiera las bazas técnicas y se descarte de las filosóficas. O las sociales y se descarte de las históricas. O cualquier otra combinación posible. Hay muchos grupos peleando sus propias partidas. Nuevas partidas añaden nuevas figuras. Y, por supuesto, siempre está la opción de romper la baraja.

Te pido que no te atasques con lo que no se entienda o no te motive. Estoy segura de que, si sigues leyendo, en algún momento encontrarás lo relevante. Darás con «tus» cartas.

Además de los naipes, en este libro hay otras cartas. Están dirigidas a unas pocas de mis amigas y amigos que están ahí siempre dispuestas a compartir ideas y hacer cosas juntas. La lectora se las puede tomar como un respiro, como una digresión o como un encuadre. Son, simplemente, trozos de conversación. Afectos.

Por qué

Carta a Marta

Querida Marta, ¿cómo va la vida?

Yo aquí, dándole vueltas al asunto. He estado pensando en lo que me propusiste, lo de participar en el encuentro de traductoras.

No tenía muy claro que pudiera aportar algo, pero… he oído un pódcast del programa *Sapiens*, el de Paula Aller en Radio Nacional. Se llama «Singularidad» y ¡uf! Resulta que el entrevistado, que es un informático profesor de inteligencia artificial en una universidad, dice que ChatGPT es una máquina que ha empezado a pensar, que abstrae conceptos complejos y que realiza razonamientos.

A ver, no es que yo quiera saber más que un profesor de universidad, pero…, ejem…, decir que ChatGPT piensa, decir que razona… me parece muy pero que muy inexacto. No sé…, yo también soy informática y no diría que las máquinas piensan. Afirmar esas cosas creo que confunde más que clarifica.

Luego habla sobre el futuro. Dice que las máquinas llegarán a tener sentimientos y consciencia. Que será posible meter en un chip todo lo que hay en la mente de un humano. Bueno, eso, como son especulaciones, pues ya lo veremos ;-)

También habla sobre los cambios en el trabajo, que ya no va a ser el centro organizador de la vida, y sobre la necesidad de desvincular actividad e ingresos, y sobre el salario social. Es interesante, pero la verdad es que la manera de explicar lo que es ChatGPT me ha inquietado bastante. Deja caer ideas que son más que discutibles. Y si la gente que las escucha no tiene criterio, pues se hará una empanada mental que no veas.

Así que sí, voy a ir al encuentro de traductoras a hablar sobre las inteligencias artificiales. A ver si entre todas digerimos la empanada. Ja, ja, ja.

Regreso pronto.

Besos.

1. Visiones

Mark I Perceptron

La máquina Mark I Perceptron se construyó en 1958 en Estados Unidos. En esa época, en plena Guerra Fría y enfrentamiento político e ideológico entre el capitalismo —liderado por Estados Unidos— y el comunismo —liderado por la Unión Soviética—, había mucho dinero para invertir en la investigación de lo que se llamaba ciencia básica, es decir, no necesariamente aplicada a soluciones concretas. La agencia de investigación militar DARPA disfrutaba de total independencia para realizar investigación no convencional. Trabajaba libre de condicionamientos, pues no se le exigían resultados rentables, y contaba con investigadores —quiero pensar que también con investigadoras— de élite que podían pensar cosas locas, aunque es de suponer que el contexto de la época, la geopolítica de bloques, tenía mucho peso.

El autómata eléctrico Mark I Perceptron estaba dedicado al reconocimiento de imágenes. Había sido entrenado con cientos de fotografías de 400 píxeles de resolución, en blanco y negro. Estas fotografías eran retratos de hombres y mujeres con distintos peinados y maquillajes, y se le había entrenado para que reconociera si cualquier rostro pertenecía a un hombre o a una mujer.

El Mark I Perceptron se había construido cableando un algoritmo, es decir, construyendo con cables una idea mecánica, que se conoce como perceptrón. Un algoritmo perceptrón sirve para realizar una clasificación binaria tipo blanco o negro, derecha o izquierda, redondo o cuadrado, etcétera. Después del entrenamiento, la máquina había aprendido a clasificar con bastante éxito los retratos de un hombre o de una mujer.

El perceptrón se concibió como una réplica del funcionamiento de las neuronas naturales, siguiendo los hallazgos de Ramón y Cajal. Es

un conjunto de neuronas hecho matemáticas capaz de aprender. Su aprendizaje consiste en que puede ir modificando progresivamente su comportamiento con el fin de acercarse lo más posible a un objetivo dado; en este caso, reconocer el género binario de un rostro en una fotografía.

Aquí, por comportamiento entendemos la respuesta que da a cada entrada, tipo 0 o 1 para cada fotografía. Y por aprendizaje entendemos el reconocimiento de patrones: el patrón hombre o el patrón mujer.

El Mark I Perceptron era un fotoperceptrón estructurado en tres capas: una capa sensorial para la entrada, una capa intermedia para el procesamiento de los patrones y una capa de resultados para ofrecer la salida.

La capa sensorial (S-units) estaba compuesta por células fotovoltaicas de dos tipos: excitadoras e inhibidoras. La capa intermedia (A-units), de asociación, era una unidad de cálculo encargada de reconocer patrones. Recibía los estímulos que le llegaban de la capa sensorial y disparaba (o no) algunas de las neuronas. Es decir, activaba sinapsis. La capa de resultados (R-units) evaluaba las sinapsis activadas y ofrecía una salida tipo 0 o 1.

El modo de aprender, es decir, la modificación del comportamiento de la máquina, consistía en que ella misma iba modificando los pesos sinápticos de las neuronas de la capa intermedia, que en el caso del Mark I Perceptron eran potenciómetros, hasta ajustarlos a unos valores que en la mayoría de los casos permitían reconocer el patrón para el cual había sido entrenada. (Ver imagen 1).

La Mark I Perceptron no era una computadora al uso. No era un ordenador ni como los de ahora ni como los de su época. Era una máquina que introducía tres innovaciones radicales respecto a lo que venía a ser el procedimiento estándar de la computación.

En primer lugar, los datos de entrada de la máquina no los proporcionaba un operador humano —por ejemplo, mediante tarjetas perforadas o mediante un teclado—, sino que los obtenía directamente del mundo físico exterior por medio de sensores. Como la idea era generalizar este modelo de máquina, los sensores ópticos serían fotorreceptores, los sensores acústicos —para realizar reconocimiento de voz— serían fonorreceptores y así seguiría con el ánimo de replicar los sentidos humanos de la vista, el oído, etcétera.

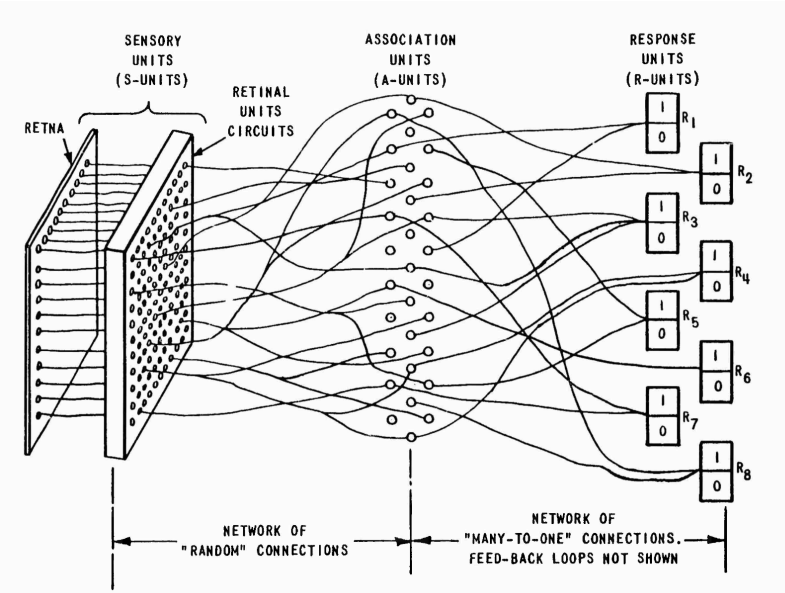


Imagen 1. El Mark I Perceptron estaba dividido en tres capas: capa sensorial (S-units), capa de asociaciones (A-units) y capa de resultados (R-units). La capa sensorial consistía en sensores fotoeléctricos que formaban una especie de pantalla de píxeles para captar la fotografía. La capa de asociaciones estaba formada por una especie de neuronas eléctricas que recibían los estímulos de la capa sensorial y, en función de cada estímulo, se activaban o se inhibían (esta era la capa de aprendizaje, la que había ajustado sus pesos sinápticos). Finalmente, la capa de resultados se encargaba de valorar todas las activaciones e inhibiciones y ofrecía un resultado binario, que en este caso era hombre o mujer. (Imagen extraída del manual de instrucciones del Mark I Perceptron, disponible en https://web.archive.org/web/20180904225539/http://www.dtic.mil/dtic/tr/fulltext/u2/236965.pdf. En YouTube se puede ver el breve documental de la época *Investigación sobre perceptrones de los años 50 y 60*, que explica su funcionamiento).

En segundo lugar, la máquina era capaz de modificar su comportamiento. Es lo que se llamó capacidad de aprendizaje. A partir de unas conexiones iniciales aleatorias, podía ir autoajustándose en función de los patrones que pudiera detectar en los datos de entrada.

Y en tercer lugar, se asumía, se daba por sentado, que el reconocimiento del patrón no lo podría realizar con precisión absoluta, sino que siempre existiría un cierto margen de error estadístico. Siempre habría rostros que reconocería mal. Siempre latiría la posibilidad de error y se daba por hecho que lo habría.

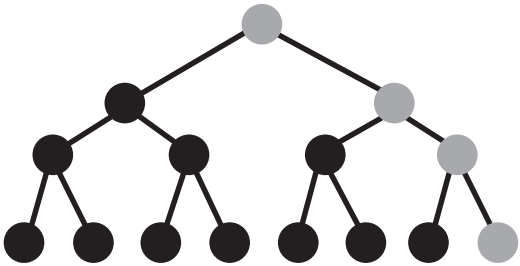
Estas tres innovaciones radicales abrían un nuevo campo de investigación aplicada: la computación de inspiración biológica. Esta rompía la senda de la computación convencional, basada en la lógica, en secuencias de pasos con una lógica interna comprensible y explicable. En cambio, la computación de inspiración biológica usaba sensores analógicos. Lógica versus analógica.

La primera procesaba símbolos en cadenas secuenciales, por ejemplo el símbolo 1 seguido del símbolo +… La segunda procesaba una amalgama de píxeles conectados todos en paralelo. Simbólica versus conexionista.

La primera era ordenada, con un orden interno comprensible y explicable. La segunda partía de un batiburrillo de conexiones aleatorias. Ordenado versus desordenado.

En una aproximación no muy rigurosa, podemos imaginar la computación simbólica como un árbol de opciones en el que la información elige un camino y excluye los demás, frente a la conexionista, que representaríamos como una maraña de opciones donde la información circula por todos los caminos a la vez. (Ver imagen 2).

En definitiva, la computación de inspiración biológica se alejaba de la lógica cristalina para enredarse en procedimientos desaliñados.



SIMBÓLICA

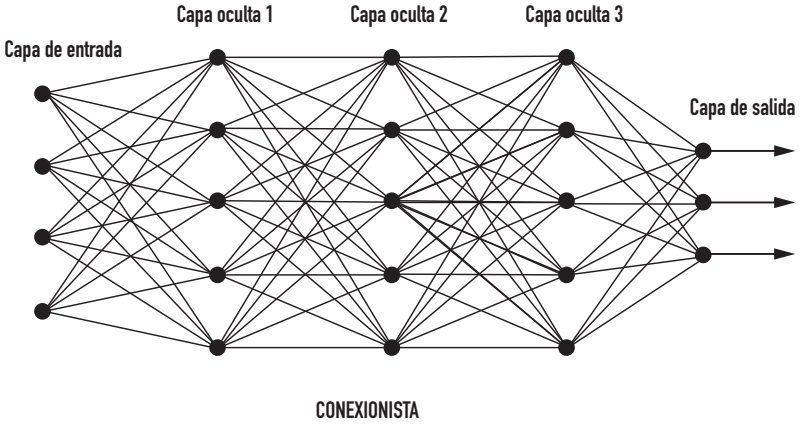


Imagen 2. Representación intuitiva de la computación simbólica (p. 28) y la conexionista (p. 29).

El Mark I Perceptron tuvo una acogida increíble. En 1958, *The New York Times* publicó nada menos que: «Es el embrión de un ordenador electrónico que espera poder caminar, hablar, ver, escribir, reproducirse y tener consciencia de su existencia». Tan alta expectativa, claro, estaba destinada a defraudar.

Desde fuera puede parecer que el ambiente científico técnico es un todo que camina al unísono, pero nada más lejos de la realidad. Las discrepancias internas son continuas y, muchas veces, enconadas. Y el perceptrón se tropezó con sus críticas.

En 1969 sus detractores publicaron estudios que demostraban que el perceptrón nunca tendría suficiente capacidad computacional para cumplir las expectativas. La manera de decirlo en lenguaje técnico es que nunca podría alcanzar la potencia de una máquina de Turing, es decir, de un ordenador convencional. Por ejemplo, nunca podría contar, así que no podría saber cuántos rostros había en una fotografía. En otras palabras, el Mark I Perceptrón se acercaba a la estructura de un cerebro, pero de un cerebro que ni tan siquiera se aproximaría a un cerebro de chorlito.

Para conseguirlo sería necesario introducir bucles, que los datos no fueran solo hacia delante, sino también hacia atrás, e introducir también más capas intermedias, más capas de asociación. Pero el problema era que no se sabía cómo programar los bucles ni el aprendizaje de las capas intermedias. Era un callejón sin salida y el asunto quedó en parada técnica.

Incompletitud

En el sigo XIX las matemáticas habían tenido un crecimiento espectacular. Teorías de conjuntos infinitos, números complejos, geometrías raras… Parte de la comunidad matemática consideraba que algunos de estos desarrollos eran demasiado abstractos e inútiles.

Las enseñanzas básicas presentan las matemáticas como algo hecho y ya está. Pero en realidad las matemáticas son un terreno vivo sujeto a oleadas que producen sacudidas y transformaciones, y siempre abierto a nuevos campos de investigación. Su interior está plagado de grandes controversias, críticas muy duras, discusiones, oposiciones, defensas enconadas y enfrentamientos exacerbados.

Hacia finales del siglo XIX se vio que algunas de estas teorías introducían paradojas que hacían que las matemáticas no fueran consistentes.

Una paradoja es algo que conduce a una contradicción y rompe la lógica. Las matemáticas se han construido desde la premisa de que no puede ser cierto algo y a la vez su contrario. Si, siguiendo sus leyes internas, se demuestra que dos más dos es igual a cuatro y a la vez se puede demostrar que dos más dos es distinto de cuatro, entonces hay un problema muy grave: las matemáticas ya no son fiables. Y deben serlo. (Aunque parezca mentira, que dos más dos es igual a cuatro se demuestra).

A finales del siglo XIX las matemáticas contenían paradojas. Sus fundamentos se tambaleaban. Eran un gigante con pies de barro.

Había que aceptarlo: las matemáticas estaban en crisis.

Una parte de la comunidad se puso manos a la obra para indagar cómo se podía resolver el asunto, cómo construir unas matemáticas libres de contradicción. No solo había que desterrar las paradojas conocidas, sino que se debía alcanzar la certeza absoluta de que en el futuro no podrían generarse otras nuevas, de que no había ninguna puerta trasera por la que se colara la contradicción.

Este programa se basaba en la convicción de que, en matemáticas, todo problema tiene una respuesta clara: o bien se demuestra su solución, o bien se demuestra que no tiene solución. No hay medias tintas. Dicho en el lenguaje de la época, se trataba de probar que el sistema matemático era completo y, además, consistente. Para ello, el camino era eliminar las intuiciones, profundizar en la lógica y apoyarse en sistemas de signos que no significasen nada concreto en relación con el mundo físico ni en relaciones y reglas de inferencia sintácticas y abstractas.

Otra parte de la comunidad expresaba sus críticas. Desde un punto de vista más intuicionista se afirmaba que las matemáticas son una actividad que tiene lugar dentro de la mente y que se expresa fuera de ella mediante el lenguaje, pero que ningún lenguaje —ni el informal ni el formal, ni el de la lógica— es lo suficientemente potente como para expresar los constructos mentales que tienen lugar en el interior del pensamiento de un humano.

Aquí la premisa era que no existe un mundo matemático fuera de la mente al que se pueda acceder mediante el uso de la razón. No hay racionalismo ni platonismo. El acto matemático emerge en la mente como un todo que explota como una erupción volcánica, todo a la vez, ya dado, y no gradualmente pasito a pasito como en una receta de cocina.

Observamos cómo en distintas filosofías de las matemáticas se planteaban distintos modos de entender qué es el lenguaje. Para las escuelas más logicistas, el lenguaje da lugar a las matemáticas. Para las más intuicionistas, lenguaje y matemáticas son fenómenos independientes. Si se les preguntase dónde reside la exactitud de las matemáticas, el intuicionista respondería que en la mente humana y el logicista que en el papel.

En estas andaba el ambiente matemático cuando en 1931 un joven estudioso de la lógica publicó el teorema de incompletitud, que demostraba que el programa logicista era inalcanzable no por falta de conocimientos, sino por límites internos esenciales: lo formal, a partir de cierto grado de complejidad, tiene fallas intrínsecas. Así como la

física se vio sacudida por la teoría de la relatividad o por la mecánica cuántica, las matemáticas se estremecieron cuando apareció ante sus ojos la incompletitud de los sistemas formales.

Un sistema formal es un conjunto de reglas determinadas que permite construir secuencias de símbolos. Podemos imaginarlo como un lenguaje con el que se generan frases. Por ejemplo, *2 + 2 = 4* es una frase de la aritmética, que es un sistema formal. Una partida de ajedrez es una larga frase generada dentro del formalismo de las reglas del juego. Tiene un aspecto similar a: *1.e4 e5 2.Cf3 Cc6 3.Ac4 Ac5 4.0-0 Cf6 5.d4 Axd4 6.Cxd4 Cxd4 7.f4 d6 8.fxe5 dxe5. 9.Ag5 Ae6 10.Ca3*… Por el contrario, el lenguaje natural, el humano, no es un sistema formal: se pueden inventar palabras, se puede retorcer la sintaxis… Un programa de ordenador es una larga frase generada dentro del formalismo del lenguaje de programación. Es algo como: *function imprimir_vars($obj) { foreach (get_object_vars($obj) as $prop => $val) { echo "\t$prop = $val\n"; } }*. Debe cumplir requisitos y reglas de sintaxis estrictas. Poner antes una instrucción que debería ir después hará que el programa no funcione como se esperaba y poner una coma donde debería ir un punto y coma convertirá el programa en algo absolutamente inservible que no se podrá ejecutar. En cambio, un texto en el que hay una frase mal colocada o con errores de puntuación, por lo general, sigue siendo muy comprensible.

Lo que demostraba el teorema de incompletitud es que todo sistema formal con una cierta complejidad nunca podrá ser al mismo tiempo consistente y completo. Si es consistente, es decir, si está libre de contradicciones, si no contiene paradojas, entonces no será completo: habrá proposiciones, frases, que dicen cosas verdaderas que no se podrán demostrar siguiendo la lógica del sistema. No se podrá definir completamente. No será completo. Tendrá agujeros. Tendrá huecos, vacíos, silencios… Su base de rigor no será total. En lenguaje matemático: habrá proposiciones indecidibles. En lenguaje filosófico: la razón no puede dar cuenta de toda la realidad. Ningún sistema racional puede comprender la totalidad de lo real.

Este teorema se mueve en un terreno muy técnico, pero nos podemos acercar a él mediante la paradoja del mentiroso.

Imaginemos un lugar donde la mitad de las personas siempre dice la verdad y la otra mitad siempre dice mentiras. Sin que sepamos a qué grupo pertenece, preguntamos a una persona: «¿Eres mentirosa?». Si nos contesta que no, su respuesta no causa ningún problema.

Si es una persona del grupo veraz, dirá la verdad («No miento») y, si es una persona del grupo mentiroso, dirá una mentira («No miento», que en este caso es mentira).

En cambio, si responde que sí, aunque en principio parezca una respuesta válida y coherente, el asunto se complica. El motivo es que, si es una persona veraz, no puede ser verdad que diga mentiras (por eso no puede decir: «Sí, digo mentiras») y, si es una persona mentirosa, sería verdad que está mintiendo y estaría diciendo la verdad, lo cual también es imposible, porque es mentirosa (no puede decir: «Sí, digo mentiras»). Estamos en una situación paradójica. Es una frase imposible. Una mentirosa que reconoce que dice mentiras es una paradoja.

Pues esto es lo que explica el teorema: que vamos a tener frases incoherentes de las que no podemos sacar nada en claro, que no podemos demostrar que sean verdaderas, pero tampoco podemos demostrar que sean falsas. En especial, todo lo que tenga que ver con la autorreferencialidad será un coladero de paradojas.

La autorreferencialidad tiene lugar cuando algo habla de sí mismo, como la mentirosa que habla de sus mentiras. En el contexto del teorema, lo que se viene a decir es que un sistema formal no puede

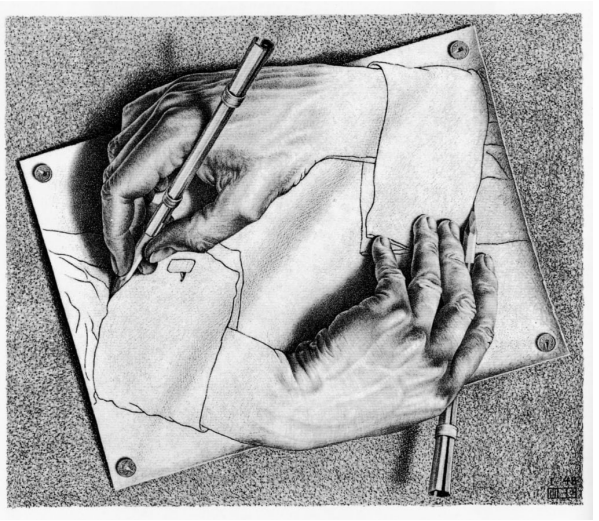


Imagen 3. En 1948 el artista neerlandés Maurits Cornelis Escher creó la litografía *Manos dibujando.* Es una situación autorreferencial imposible, paradójica.

demostrarse a sí mismo, pues al hacerlo introducirá paradojas. Para no introducirlas, tiene que haber un sistema exterior, un lenguaje exterior desde el que hablar: un lenguaje que hable de otro lenguaje, es decir, un metalenguaje.

Por ejemplo, un autorretrato no es paradójico. La pintura (el metalenguaje) habla de la persona autorretratada. La persona autorretratada utiliza un sistema exterior (la pintura) para describirse a sí misma.

Cuando algo habla de sí mismo en el mismo nivel, dentro del mismo sistema formal, entra la paradoja. Para esquivarla, tiene que haber dos niveles: el lenguaje que habla (metalenguaje, metasistema) y aquello de lo que se habla. Un sistema no se puede demostrar a sí mismo. Para demostrarlo completamente hay que poder salir fuera de él. Hay que hablar de él desde fuera. (Ver imagen 3).

Así quedaba eliminada de un plumazo la pretensión de alcanzar un conocimiento completo y cierto (insistimos, dentro de un sistema formal con una cierta complejidad, como la aritmética), porque en algunos puntos el sistema formal es ciego a la verdad. ¡Menudo bajón! Lo verdadero y lo demostrable se separaban. Y el mundo se tambaleó.

Fuera de las matemáticas, el teorema salpicó las disciplinas sociales. Julia Kristeva lo llevó a la semiótica; Deleuze, Guattari, Lyotard o Lacan se lo llevaron, cada cual, a su terreno. Conectó con todo lo que tuviera relación con información o símbolos. Y, por supuesto, impregnó las ciencias de la mente y la naciente ciencia de la computación, que se estaba preguntando qué tipo de problemas tenían solución algorítmica y cuáles no: qué era lo que una computadora podía resolver sin entrar en bucle.

Parecía demostrado que el teorema establecía la imposibilidad de que un programa de ordenador pudiera rivalizar con la mente humana, porque un programa es un sistema formal y, por tanto, está sujeto a los límites que señala el teorema. La mente humana es superior, porque puede identificar lo verdadero, aunque dentro del sistema formal no pueda demostrarlo. Así, podemos decir: «Sé que eso es verdad, aunque con las reglas formales de ese sistema no puedo demostrarlo».

Además, del teorema se deduce que no podemos hacer un programa de ordenador para verificar la corrección de otro programa de ordenador. La corrección de un programa no se puede verificar dentro de las reglas de la programación, aunque sí se podría

verificar desde fuera, desde una instancia exterior a esas reglas: la mente humana. Hay tareas que la mente humana puede realizar y una computadora no.

Entonces ¿qué pasa con el Mark I Perceptron?

Mientras que en una esquina había personas estudiosas de la lógica enfadadas, consternadas o deprimidas, en la otra punta resonaban voces que decían: «Eh, gente, dejémonos de rollos y ¡computemos!».

Aunque se había abierto una crisis conceptual, las calculadoras seguían calculando, las geometrías seguían geometrizando, todos los teoremas demostrados seguían siendo demostrables… El día a día matemático seguía su curso, igual que sigue mi día a día aunque no entienda nada de la teoría de la relatividad o de la mecánica cuántica. Al fin y la cabo, las matemáticas son un montaje intelectual y tienen la trascendencia que les quieras dar. Es verdad que se habían señalado unos límites, pero todavía quedaba mucho recorrido.

Sin embargo, la gente que decía «¡Computemos!» no estaba en ese punto, sino en otro mucho más allá. Negaban la mayor. ¿Quién ha dicho que las mentes humanas son completas? El teorema también se puede aplicar a la mente humana, así que una mente no puede pretender ser superior a una máquina. De hecho, por el mismo argumento, una máquina puede defender su superioridad sobre una mente, ya que el teorema se aplica tanto a las máquinas como a las mentes. En resumen, en lo que se refiere al pensamiento, la mente humana es mecanizable.

Estaban en disposición de abrir nuevas vías de investigación en lógica y en el conocimiento de la capacidad cognitiva e intuitiva de la mente, en la diferencia entre tener una idea y seguir un procedimiento mecánico o algorítmico. Querían entender cómo funcionaba la creatividad matemática humana y llegar a los límites de lo cognoscible. Sospechaban de la propia consistencia de la mente, de que en su interior no albergue la contradicción. Querían entender los límites de lo humano y daban la vuelta al calcetín: quizás nuestros razonamientos sean válidos, pero no podemos estar seguros de ello. Si no puedes demostrarlo, no puedes tener certezas. Nuestro pensamiento tiene límites, muchos límites. Así que los procesos mentales no pueden ir más allá que los maquínicos. Lo humano y lo maquínico responden al mismo modelo.

Kybernetes

Durante la Segunda Guerra Mundial, la ingeniería de sistemas de control encontró un asunto interesante en el problema del control de tiro. Con la guerra, en Estados Unidos la industria, el gobierno y las universidades se articularon para cooperar en torno a la investigación aplicada a la defensa militar. En el National Defense Research Committee (NDRC) se investigaba sobre la bomba atómica y el radar, sobre el desarrollo de modelos matemáticos para simular la toma de decisiones y también sobre el control de tiro en la artillería antiaérea, las ametralladoras que disparan proyectiles contra aeronaves en vuelo cuando ataca la aviación enemiga. La cuestión era cómo usar el radar para orientar los proyectiles contra un objetivo móvil.

Desde el punto de vista matemático, se vio que para guiar con precisión el proyectil hacia el objetivo había que predecir el futuro de una curva (la curva de la trayectoria de la aeronave en movimiento a gran velocidad) y predecir el futuro de una curva implicaba tener en cuenta su pasado (la trayectoria ya realizada). En relación con este guiar, surgió una idea novedosa: la máquina que se controla a sí misma.

Es algo parecido a cuando tengo que regular el agua caliente de la ducha en un hotel con un sistema de grifos que no conozco. Además de poner la mano para captar la temperatura del agua, tengo en cuenta cuánto he girado la maneta del monomando. El giro que voy a hacer, que espero que sea certero, tiene en cuenta el que he hecho antes en combinación con la temperatura que he conseguido. Lo que hago depende de lo que he hecho antes. Así, aprendo cómo funciona ese sistema de grifos y voy reajustando su apertura hasta alcanzar la temperatura deseada. Si no tuviera memoria del movimiento que he

hecho antes, seguramente no lo conseguiría. Aun así, a veces cuesta bastante.

Siguiendo este modelo, la máquina autocontrolada es un sistema que se controla a sí mismo mediante el uso de la retroalimentación, y es capaz de adaptarse y ajustarse a los cambios de su entorno sin necesidad de intervención humana constante. La máquina puede ser programada para regular su propio comportamiento si se le enchufa retroalimentación. Es un servomecanismo. Una máquina con intención.

La retroalimentación es un proceso dinámico por el cual la información de salida de un sistema se redirige para que vuelva a entrar en el mismo como información de entrada. En cada momento, la máquina tiene información de lo que ha hecho justo en el momento anterior. De esta manera tiene capacidad autónoma para autorregularse ella misma a partir de sus propios errores, con el fin de alcanzar un objetivo predefinido. De alguna manera, sabe por dónde va y adónde tiene que ir, así que puede modificar continuamente su comportamiento para orientarlo hacia la meta preestablecida. Con las máquinas autocontroladas surge la cibernética.

Un ejemplo intuitivo de máquina cibernética es la cisterna del inodoro. Cuando se vacía, el propio acto mecánico del vaciado baja la boya y esto abre el grifo para que se vuelva a llenar. Mientras se llena de agua va subiendo la boya, hasta que alcanza un nivel determinado que cierra el grifo. El propio nivel del agua es el que abre o cierra el grifo. El agua tiene información sobre sí misma, sobre su nivel, y abre o cierra el mecanismo para autoajustarse al objetivo de mantenerse en un nivel. El agua se regula ella misma (sirviéndose de la boya).

«Cibernética» ahora es una palabra sin *glamour* que evoca el mundo gris de la posguerra, un mundo de masas productivo e industrial. Sin embargo, no era esa la percepción que se tenía de ella en los orígenes de esta ciencia que creció en los espacios liminales, en la tierra de nadie que quedaba en medio de las ciencias ya constituidas.

El término hace referencia al *kybernetes* griego, que para Platón era el arte de pilotar una nave. Mientras que el capitán fija el destino y el timonel maneja el timón, el piloto marca el rumbo, un rumbo que reajusta constantemente teniendo en cuenta su conocimiento sobre la nave y leyendo e interpretando los continuos cambios en los vientos, el estado de la mar… Los que actúan deben considerar

siempre lo que es oportuno, como ocurre en el arte de la medicina y en el del pilotaje, dice Aristóteles, que asocia el término *kybernetes* a la acción prudente. No es tanto mandar como tener en cuenta los hechos, saber leer sus mensajes.

El problema del control de tiro, además de inspirar la construcción de servomecanismos con capacidad de autocontrol, planteaba otra cuestión si cabe aún más interesante. El sistema tenía componentes operados por humanos: el artillero y el piloto del avión objetivo. Esto hizo que algunos matemáticos señalaran que había que centrar la atención tanto en los dispositivos maquínicos como en el comportamiento de los humanos que los manejaban. Específicamente, había que centrarse en el acoplamiento hombre-máquina.

Separándose del conductismo, se señaló que era necesario el estudio de los procesos mentales implicados en la toma de decisiones. Había que estudiar al operador humano como un componente integrado de un sistema de control automático. Había que entender cómo producir un ensamblaje óptimo en la interfaz humano-máquina, cómo conseguir un ajuste perfecto entre las capacidades del operador y las de la máquina. Había que saber cómo el factor humano contribuía a la estabilidad del sistema o bien introducía inestabilidad emocional, especialmente cuando el humano operaba bajo fuego enemigo, ya que este conocimiento era determinante para solucionar el problema del control de tiro. Es así como la cibernética se constituye como ciencia transversal que utiliza las matemáticas, la lógica y las ingenierías, pero también la biología y la psicología.

Desde esta mirada transversal, veían que el modelo del bucle de retroalimentación era aplicable a una gran cantidad de fenómenos que abarcaban el funcionamiento de los seres vivos, los cuales también regulan su entropía mediante la retroalimentación. La cibernética explica tanto el comportamiento fisiológico como el de las máquinas. Entre ambos hay un paralelismo exacto. El piloto de un avión se comporta como un servomecanismo. La lógica humana y la lógica de la máquina son lo mismo.

Esta nueva ciencia generaba tales expectativas que desde la sociología y la economía se le pedía a la cibernética que propusiera soluciones a los problemas sociales de la época. Desde la antropología, figuras como Margaret Mead llamaban a movilizar conocimientos y recursos de la cibernética para ver cómo se podían paliar «los problemas sociológicos y económicos de la presente era de confusión».

Sin embargo, desde la propia cibernética se tomaba distancia: las matemáticas no se pueden aplicar así por las buenas a las ciencias sociales, porque las ciencias sociales no se pueden aislar en un laboratorio que estabilice todas sus variables. Las ciencias sociales no son un buen campo de demostración de nada. Son demasiado indefinidas como para hacer probaturas con ellas. El entorno social no se puede controlar. Donde sí se pueden probar estas ideas es en la ingeniería y en la biología, en los organismos vivos y en las máquinas.

A principios de los años setenta, el Chile de Allende fue uno de esos escenarios en los que la utopía cibernética se puso manos a la obra para experimentar en un proyecto audaz: el Cybersyn, también conocido como Synco.

El objetivo de este proyecto era reorganizar la producción de las numerosas empresas nacionalizadas. Cada fábrica nacionalizada, a través de una red de comunicaciones, enviaría en tiempo real datos sobre su producción a una especie de sala de operaciones que centralizaría la información. Un equipo humano elegido democráticamente podría realizar cálculos y tomar decisiones sobre la producción que a su vez llegarían a las fábricas en tiempo real a través de la misma red. Así, las preguntas sobre si se estaba produciendo suficiente comida o demasiados vehículos tendrían respuesta.

Ese era el objetivo, pero el sentido de Cybersyn era orientar la estructura productiva en beneficio de la población bajo un modelo socialista que se alejara tanto del libre mercado capitalista como de la aplastante planificación soviética. Control y comunicación en estado puro. Acoplamiento entre humanos y máquinas. En un Chile en el que había menos de cincuenta computadoras, ideas vanguardistas gestadas en debates universitarios, comités políticos a favor de la justicia social y procesos sociales que luchaban por la soberanía popular crearon las condiciones para que una generación de jóvenes soñara con una cibernética para el cambio social y se sintiera capaz de construirla.

No iba a ser tan sencillo, porque el golpe militar de 1973 acabó con el proyecto; pero Cybersyn no murió sin pena ni gloria. En 1972, cuando todavía se encontraba en estado de prototipo, tuvo lugar el paro de los camioneros. Fue un paro patronal apoyado por la CIA para desestabilizar al gobierno socialista en el que participaron más de 8.000 camioneros. Con carreteras bloqueadas y sin camiones para suministrar alimentos ni materias primas, gobierno y trabajadoras se

autoorganizaron usando Cybersyn para intercambiar información clave y montar redes de suministro con las que gestionar la escasez. No estamos hablando (solo) de ingenieros o de matemáticos. Las organizaciones obreras y populares, grupos de barrio, de mujeres… consiguieron mantener la producción y los suministros en niveles suficientes como para que el paro de camioneros perdiera eficacia. La utopía funcionó.

La cibernética no es popular. Es como si hubiera pasado a la clandestinidad. No es sinónimo de robótica ni de informática, ni tampoco de automatización. Es un pensamiento transdisciplinar que propone soluciones creando sistemas que implican a todas las partes. Su punto fuerte es activar las relaciones, construir puentes que realimentan las relaciones.

La cibernética es la ciencia del control y la comunicación en animales y máquinas. Es la ciencia de los organismos. Su tesis es que animal y máquina, independientemente de que su naturaleza física sean proteínas o circuitos electrónicos, se pueden explicar por un mismo dispositivo formal: la retroalimentación. Hay un modo homogéneo de explicar la conducta de los animales y de las máquinas. En el acoplamiento animal-máquina hay una total continuidad. Entre animal y máquina no hay ruptura.

Desde su comprensión del funcionamiento de los sistemas de control y comunicación en organismos vivos y máquinas, desde su explicación de la naturaleza de la realidad, no hay ningún problema en decir, metafóricamente, que una máquina tiene memoria o que aprende, ya que esto simplemente son atajos, sentidos figurados más o menos útiles. La cibernética no dice que la máquina tenga atributos humanos. Lo que dice es que el humano es asimilable a la máquina. Y en esta asimilación la máquina es el modelo.

Como destilación de este paradigma, una parte de la cibernética se podrá sobreponer al profundo pesimismo que, dentro de sus propias filas, causa la mecanización de lo humano con la propuesta alegre y esperanzadora de la humanización de la máquina: un acoplamiento entre humano y máquina en continuidad, en horizontalidad, en autoorganización sin mando, y siempre con el reconocimiento y el respeto a un tercer término: el mundo en el que ese acoplamiento tiene lugar.

Las máquinas son parte de la realidad humana, son gesto humano fijado y cristalizado en estructuras que funcionan. No estamos por encima de ellas. Estamos con ellas. Entre ellas. Son nuestros iguales.

Explícame

En los años setenta ya había modelos matemáticos de redes neuronales artificiales capaces de aprendizaje automático, como por ejemplo el perceptrón. Sin embargo, no estaba muy claro cómo ponerlos a funcionar, así que una parte de la investigación optó por un cambio de estrategia y tiró por otro lado: volvamos a la lógica.

El objetivo era desarrollar programas informáticos que aplicaran la computación cognitiva a problemas de la vida cotidiana y pudieran imitar las capacidades de toma de decisiones de una persona experta en un área de conocimiento acotada. Por ejemplo, un sistema médico que pudiera diagnosticar la enfermedad de un paciente basándose en los síntomas y en el historial clínico. Esos programas se llamaron, cómo no, sistemas expertos.

La medicina se presentaba como uno de los campos de aplicación para los sistemas expertos. El sistema podría servir de ayuda a profesionales no especializados, compensando así la distribución geográfica irregular de los doctores y doctoras especialistas. Se trataba, por tanto, de construir sistemas expertos muy especializados enfocados, cada uno de ellos, en un área de conocimiento muy específica. MYCIN fue uno de esos sistemas.

MYCIN era un sistema experto médico especializado en diagnosticar enfermedades infecciosas en la sangre producidas por bacterias, estimar su gravedad y proponer un tratamiento con antibióticos concretos y con la dosis adecuada para cada paciente. Ante una infección en la sangre, se pueden realizar extracciones y enviarlas a un laboratorio, pero ese proceso en el mejor de los casos llevará uno o dos días y es muy arriesgado tener a un paciente infectado tantas horas

sin tratamiento, por lo que el doctor o doctora tiene que diagnosticar basándose en síntomas. En ese contexto, un sistema experto de ayuda en la toma de decisiones era un opción muy útil. MYCIN podía detectar la meningitis y la bacteriemia. Se bautizó así por su similitud con el nombre de los antibióticos que prescribía.

Los sistemas expertos, tal como se abordaban en los años setenta, suponían un regreso a la computación convencional basada en la lógica, en secuencias de pasos racionales y deductivos con una coherencia interna comprensible y explicable. Había que modelar una base de conocimiento humano y había que hacerlo con orden y estructura, es decir, moviéndose en un terreno simbólico que estaba en las antípodas del conexionismo del Mark I Perceptron.

Había que usar la lógica, sí, pero ¿cuál? Una candidata era la lógica proposicional, pero resulta que no era lo suficientemente potente, porque no permite distinguir entre el sujeto y el predicado. Por ejemplo, en este acertijo lógico que dice: «El asesino de Roberto estuvo en la habitación del hotel», «La señora García no estuvo en la habitación del hotel», «Luego la señora García no es la asesina de Roberto», está claro que la deducción es evidente y verdadera. Pero, si se tienen que procesar las frases como un todo y no hay manera de descomponerlas separando el nombre de la persona (sujeto) de la acción de estar en la habitación del hotel (predicado), entonces no hay una manera mecánica, algorítmica, de llegar a la deducción.

Por tanto, había que tirar de una lógica más compleja: la lógica de predicados de primer orden, que permite entrar en el contenido de las frases, descomponerlas y manejar relaciones, por ejemplo la relación «estar en la habitación del hotel». Pero al aumentar la complejidad de la lógica la complejidad computacional aumentaba de tal manera que la programación de los sistemas expertos se hacía imposible de todas todas, porque la complejidad subía hasta un escalón muy alto, que en lenguaje técnico se llama NP-completo.

Todo programa de ordenador, para ejecutarse, necesita dos recursos: una memoria y un procesador. La memoria aporta un espacio más o menos grande para guardar los datos que se están procesando u otros que pueden ser de utilidad en el proceso. Y el procesador aporta la capacidad de cálculo que permite procesar los datos. Con el tiempo, las memorias se han hecho más grandes y los procesadores más rápidos, de modo que muchos algoritmos que antes no se podían

ejecutar por falta de recursos ahora sí son ejecutables. Muchos, pero no todos.

Desde los inicios de la computación surgió la inquietud por establecer qué era lo que se podía resolver mediante un algoritmo y qué no. Se quería estudiar la complejidad algorítmica y calcularla exactamente para que, teniendo dos algoritmos que resolvieran el mismo problema, se pudiera optar por el más eficiente, es decir, el menos complejo, el más rápido. En el lenguaje informal la complejidad es algo difuso, pero en computación es una medida exacta, un número que cuenta la cantidad de pasos que el procesador tiene que dar para ejecutar el algoritmo, en función del tamaño de los datos a procesar. Cada algoritmo tiene una complejidad exacta, un numerito que se calcula atendiendo a su lógica interna y que varía en función del tamaño de los datos a procesar.

Por ejemplo, para ordenar alfabéticamente una lista de palabras existen muchos algoritmos de distintas complejidades. Si se elige el que se conoce con el nombre de algoritmo de la burbuja, se sabe que su complejidad es $n(n-1)/2$. Esto quiere decir que, si se ordenan 10 palabras, la cantidad de pasos que tiene que dar el procesador es de $10*9/2 = 45$. Y, si tienen que ordenar 217 palabras, la cantidad de pasos será de $217*216/2 = 23.436$. (La n es el número de palabras a ordenar).

Pues bien, estudiando las complejidades se llegó a la conclusión de que hay un tipo de problemas en los que, aunque haya algoritmos que los resuelvan (son computables), la complejidad, es decir, el número de pasos que hay que dar para resolverlos y, por tanto, el tiempo que se va a emplear en ello, a poco que la cantidad de datos a procesar sea mínimamente grande, crece tanto que es como si se fuera acercando al infinito. Tiende al infinito. Esta clase de problemas, o de algoritmos que los resuelven, se denominan NP-completos. Son problemas que en teoría se podrían resolver, porque se conocen algoritmos que los resuelven, pero para ello se necesitaría un tiempo muy grande, cercano al infinito, así que a efectos prácticos son inmanejables. En lenguaje técnico se llaman intratables. Y la solución de algunos problemas con lógica de predicados de primer orden entraba ahí.

NP-completo es inabordable, intratable, porque la cantidad de pasos enseguida sube hacia el infinito y, por más velocidad que alcancen los procesadores, al infinito nunca se llega. (Ver imagen 4, en p. 46).

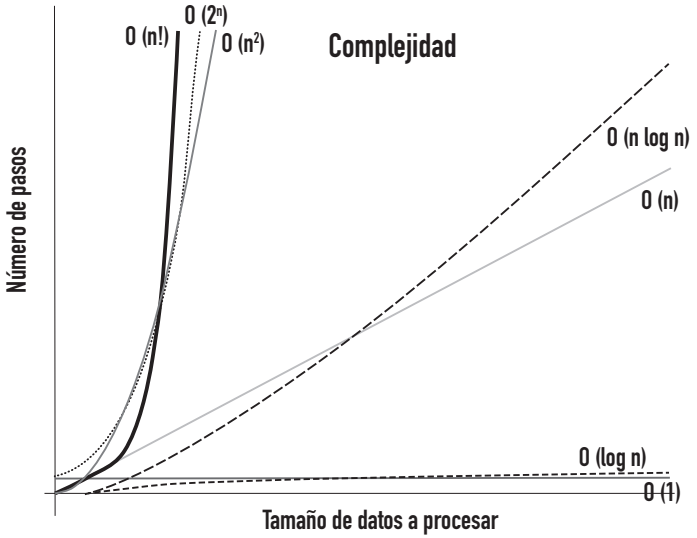


Imagen 4. En la imagen, la línea O (n!) muestra el crecimiento de la cantidad de pasos necesarios para resolver un problema de complejidad NP-completa en función del tamaño de los datos (en el eje horizontal). Es la de mayor crecimiento.

Así que había que hacer algo para reducir la complejidad y se ideó un camino: en lugar de usar toda la lógica, hagamos sistemas basados en reglas, que son más sencillos y pueden funcionar. Quitemos la combinatoria, que es lo que hace que la complejidad sea intratable, tracemos caminos, itinerarios deductivos transitables, y podemos el resto de ramificaciones. Construyamos reglas que dibujen un mapa dentro del laberinto. (Ver imagen 5, p. 47).

¿Cómo construir esas reglas? ¿Cómo saber qué caminos hay que podar? Pues consultando a las personas expertas, profesionales de la medicina especialistas en infecciones en la sangre. Para la primera versión de MYCIN, después de consultar con expertos, se construyeron unas doscientas reglas, que se fueron ampliando hasta más de quinientas en las siguientes versiones. Esas reglas expresaban el conocimiento médico en esta área de diagnóstico.

El programa preguntaba a las pacientes sobre sus síntomas. Eran preguntas con respuesta de sí o no. A partir de las respuestas hacía un diagnóstico y proponía un tratamiento. Funcionaba muy bien, salvo

por un problema que siempre se cuela en los sistemas lógicos: no se podía tratar con la incertidumbre. El sistema sabía cómo tratar si a la paciente le dolía la cabeza o no le dolía, pero no sabía cómo tratar el hecho de que en algunos momentos le doliera un poco. La lógica no se maneja bien con la incertidumbre.

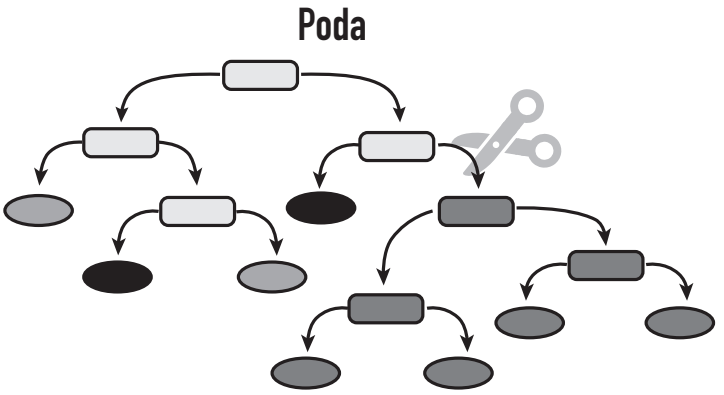


Imagen 5. Para reducir la complejidad, se podan (se eliminan) caminos.

¡Que no cunda el pánico! Podemos resolverlo, inventaremos nuevas lógicas, enseñaremos a la lógica a manejar la incertidumbre. Así, las lógicas existentes se extendieron con nuevos desarrollos matemáticos que inventaron otras lógicas: las modales, las difusas…, pero no terminaban de funcionar.

La cuestión es que no se podía usar la probabilidad, que era la herramienta conocida más potente para tratar la incertidumbre. Y no se podía usar porque, si se usaba, se volvía a introducir la complejidad NP-completa. Los sistemas expertos no podían funcionar sobre la base de probabilidades, porque eso era de complejidad intratable; pero había más estrategias. En lugar de pedir al algoritmo que realice un cálculo de probabilidades a lo bestia, vamos a usar una probabilidad acotada: pediremos a las personas expertas que indiquen la probabilidad de que ocurra una cosa si ya ha ocurrido otra. Las personas expertas, por su experiencia, pueden aseverar que, si a ratos te duele un poco la cabeza y además tienes tanto de fiebre, entonces tienes una probabilidad de tal tanto por ciento de estar padeciendo tal

enfermedad. También es probabilidad, pero probabilidad ya calculada por humanos.

Y así se hizo. Se diseñaron modelos basados en factores de certeza y todo parecía funcionar muy bien. El MYCIN tenía una base de conocimientos con las reglas y los factores de certeza; tenía un motor de inferencia que, a partir de los datos que entraba la persona usuaria y en combinación con la base de conocimientos, llegaba a una conclusión: hacía un diagnóstico, pronosticaba la gravedad y prescribía un tratamiento. Era un sistema determinista: con las mismas entradas daba las mismas salidas. El azar no tenía cabida.

Se hicieron pruebas para evaluar la calidad del sistema y se vio que MYCIN realizaba mejores diagnósticos que los doctores o doctoras no especialistas en ese tipo de infecciones, y eso se consideró muy exitoso y útil.

Además, tenía muchas ventajas. Una de ellas es que se podían ir añadiendo nuevas reglas sin tener que rehacerlo todo, así que se podía mejorar sin mucho esfuerzo. Pero todavía tenía algo mejor: era capaz de dar explicaciones, era un sistema explicable. Se le podía pedir que justificase sus recomendaciones, que explicase el razonamiento que había seguido, y ¡podía hacerlo! Claro, era un sistema lógico y, por tanto, era explicable. MYCIN no era una caja negra. Estaba hecho de pura lógica y daba explicaciones. Y eso todavía lo hacía más útil.

En computación se habla de cajas negras en un sentido distinto al de las cajas negras de los aviones. La caja negra de un avión es un dispositivo que durante la navegación graba y almacena todos los datos importantes (altitud, velocidad, rumbo…), así como las conversaciones en la cabina, etcétera. Su función es saber qué ha pasado. En computación, una caja negra es todo lo contrario. Se dice que un sistema es una caja negra cuando se desconoce su funcionamiento o cuando el funcionamiento, visto desde fuera, no se comprende. Una caja negra es un dispositivo o un sistema que oculta sus detalles internos.

Es cierto que, desde el punto de vista de la consistencia, la gente que estaba a cargo de la programación advertía de que cuando las personas expertas ponderaban los factores de certeza se introducía subjetividad. De hecho, dos expertas podían tener opiniones contradictorias. Así ocurre en la medicina real. Pero no fue ese el motivo por el que MYCIN nunca se puso en funcionamiento.

Había un problema mayor y no era de índole técnica, sino social. Si los sistemas sanitarios no se atrevieron a implantarlo no fue por falta de fiabilidad, sino por un asunto ético o legal: si el programa realizaba un diagnóstico erróneo con consecuencias negativas para la paciente, incluida la posibilidad de muerte, ¿quién sería responsable? ¿El facultativo que lo había usado como soporte para su decisión? ¿Los doctores o doctoras que habían definido las reglas? ¿Las personas que habían ponderado los factores de certeza? ¿Los programadores?

El entorno social en el que MYCIN se tenía que acoplar no lo vio claro. Había preguntas. Y MYCIN nunca se usó.

Trabajar

Carta a Jabuti

Querido Jabuti, ¿qué tal vais?

El otro día te echamos de menos en la reu de traductoras. Estuvo muy bien. Salieron muchas cosas de cómo los traductores automáticos están cambiando el trabajo, de la precarización… Había preocupación y susto.

Me he leído la declaración de UniCo, esa que me mandaste sobre el uso indiscriminado de la inteligencia artificial generativa en la corrección. Claro, para las correctoras también está cambiando el mercado de trabajo. Todas las empresas quieren ahorrar y cepillarse a la correctora parece la primera opción. A este paso, no sé yo si llegarás a la jubilación. Ja, ja, ja.

Lo interesante de vuestra profesión es que, como dicen en la declaración, la mayoría de las propuestas de regulación se centran en que esos sistemas se están entrenando con contenidos que tienen derechos de explotación, pero no tienen en cuenta a otra gente que también se precariza y que no tiene esos derechos.

Bueno, a ver si a fuerza de reunirse la gente salen reivindicaciones que se puedan pelear.

Y cambiando de tema, ¿qué tal las extraescolares de ajedrez? Una pena que te cueste retener a las niñas, snif.

Yo estoy haciendo lo que me dijiste al pie de la letra. En Lichess no jugar contra la máquina, jugar solo contra humanos. Y siempre analizar la partida. Lo que pasa es que, cuando el motor de ajedrez me da el análisis y me dice los errores, muchas veces no entiendo por qué eso es un error. Agggggh. Da rabia saber algo, pero no saber el porqué. Disfruto más entendiendo la partida que ganándola. Poco competitiva que es una… :-)

Por cierto, vete reservando hueco para corregir el libro. He decidido usar el femenino genérico, todas somos las personas, salvo en casos muy concretos, je, je. También he

decidido no poner notas al pie, porque ahora con buscar en Internet es muy fácil encontrar las cosas y este libro tampoco es una gran investigación. A ver qué dice a eso tu criterio como corrector. Un amigo me ha dicho que eso mismo es lo que hace ChatGPT, no citar sus fuentes. Ja, ja, ja. Bueno, asumo la paradoja ;-)

Y se me quedó rebotando en la cabeza lo que contaste de Marrakech: que antes los turistas no se atrevían a entrar en algunas zonas de la medina porque tenían miedo de desorientarse, perderse, no encontrar la salida y meterse en líos; pero ahora con el Google Maps no hay quien los pare, se atreven a meterse por todos lados. Es como si la gente de la medina hubiera estado protegida de la invasión turística por su saber orientarse dentro del laberinto y Google Maps les hubiera arrebatado ese saber. Ahora se pueden pisotear las calles de la medina de cualquier manera solo para obtener un rato de experiencia exótica. Ahora cualquiera sabe orientarse en esos espacios antes negados a los intrusos. Ejem, «saber» es un decir. Lo que cualquiera sabe es caminar mirando el móvil ;-)

Hablando de caminar, cuando tengas tiempo me avisas y nos damos un paseo.

Besos.

2. Espectacular

Enséñame

Desde siempre, la computación se ha interesado por el ajedrez y el ajedrez se ha interesado por la computación. Este deporte mental, del que se dice que sobre el tablero hay más posiciones posibles que átomos en el universo, despliega un estimulante campo de batalla en el que la evaluación de esas posiciones marca el plan estratégico que conducirá a la victoria o a la derrota en la partida.

Desde la programación lógica, el universo estructurado y reglamentado del ajedrez ofrece un terreno de juego tentador para probar la potencia de cálculo de los circuitos electrónicos y la potencia deductiva de la lógica computacional y de los algoritmos. Así que la lógica del juego y la lógica computacional se aliaron para crear los motores de ajedrez.

Un motor de ajedrez es un programa de ordenador que juega al ajedrez y se puede comunicar con humanos o con otros motores. Es un sistema experto en ajedrez.

Las personas que juegan al ajedrez tienen que comparar su nivel de juego respecto a otras personas, así que hay sistemas de puntuación que miden la fuerza relativa de una jugadora en comparación con otras. El sistema de puntuación más utilizado y reconocido por las federaciones es el ELO.

El ELO es un número que se asigna a cada ajedrecista y se calcula según la evaluación de las partidas que ha jugado antes. Después de cada partida jugada, el ELO se actualiza y sube o baja en función del resultado de esa partida, así que es un número dinámico. Los motores de ajedrez también son ajedrecistas y también tienen ELO. En el momento

de escribir estas líneas, mi entrenador de ajedrez tiene un ELO de 1.717; el del actual campeón mundial, Magnus Carlsen, es 2.833; y el del motor de ajedrez Stockfish es 3.546. Claramente muy superior.

Stockfish es un motor de ajedrez que puede correr en cualquier ordenador personal o teléfono con Android o iOS. Es gratuito y se distribuye como software libre con la licencia GPLv3. Eso significa que su código está disponible en GitHub y cualquier persona lo puede leer, descargar y modificar. Está desarrollado y mantenido por una comunidad colaborativa, en este momento formada por 241 personas voluntarias. Las personas que no son programadoras también pueden colaborar ofreciendo la capacidad de cómputo de su ordenador para probar las futuras versiones que se están desarrollando. Eso se puede hacer sin tener ningún conocimiento de programación. Y también se puede participar en los debates que los desarrolladores mantienen en los foros de Discord.

En una partida de ajedrez, la magia que conduce a la victoria es saber decidir, cuando toca el turno de mover, cuál jugada, de entre todas las posibles, es la mejor. Es cuestión de cálculo, pero de un cálculo endiablado. Nada más empezar, después del segundo movimiento de las blancas, puede haber 5.326 posiciones de tablero distintas. En la posición en la que se encuentre el tablero, la jugadora que tiene el turno tiene que evaluar cuál, de todos los posibles movimientos que puede hacer, es el mejor. La jugadora no puede evaluar todas las opciones, porque le estallaría la cabeza. Examina solo unas pocas: las que cree que tienen probabilidad de éxito. Mejorar el juego consiste en ir identificando las buenas elecciones, aprender a separar el trigo de la paja, y hacerlo con agilidad para no perder la partida por tiempo.

Para un programa de ordenador, el problema es distinto, porque tiene mucha capacidad de cálculo y mucha velocidad, y puede permitirse el lujo de evaluar más opciones que el humano. En lenguaje técnico, se dice que tiene más fuerza bruta. Stockfish es un motor de ajedrez de fuerza bruta.

Para determinar la mejor jugada, el mejor movimiento, un motor de fuerza bruta utiliza una combinación de algoritmos de búsqueda y evaluación de posiciones. El algoritmo de búsqueda sirve para hacer una lista de todos los movimientos posibles. Luego realiza cada uno de esos movimientos uno por uno y evalúa las posiciones resultantes. Es decir, empieza por el primer movimiento de la lista, lo realiza y evalúa la posición resultante. Después lo deshace, realiza el segundo

movimiento de la lista y vuelve a evaluar la posición resultante. Y así con toda la lista. Al final, ordena todas las evaluaciones y elige la de valor más alto. Ese será el mejor movimiento.

En realidad, si lo hiciera de esta manera, sería un motor de profundidad uno. Pero los motores tienen una profundidad de veinte o incluso más. Eso significa que, para cada movimiento de la lista, repite los pasos hasta veinte veces o más: realiza el movimiento y lo evalúa, a continuación evalúa, en esa posición, el mejor movimiento de la contrincante (lo cual requiere repetir todo el proceso de hacer la lista, etcétera), y así avanza hasta veinte movimientos consecutivos, con lo que alcanza una profundidad de veinte, es decir, ve cómo quedará la posición dentro de veinte movimientos si ambos contrincantes juegan sucesivamente los mejores de los posibles.

Lo que llamamos buscar, hacer una lista, realizar un movimiento, evaluarlo, ordenar, etcétera, para el motor consiste en manejar modelos de datos y realizar cálculos con ellos. Es decir, tener una representación matemática de la realidad que quiere representar, el juego de ajedrez, y realizar acciones matemáticas manipulando esos datos.

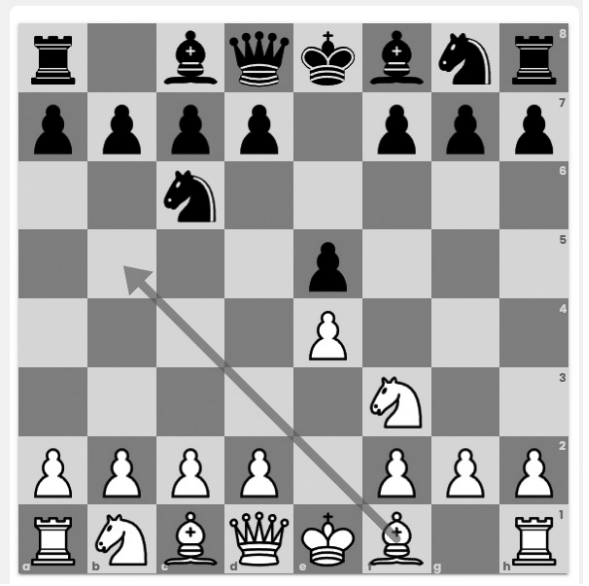


Imagen 6. Ante una determinada posición de las piezas sobre el tablero, el motor propone con una flecha la mejor jugada posible. (Ver también imagen 7, p. 60).

La calidad de un motor de ajedrez viene dada por dos características. Una es la velocidad. Para ganar velocidad, en lugar de evaluar todas las posiciones de la lista, el motor debe descartar (podar) las que es evidente que no son buenas. Por ejemplo, las que colocan una pieza en una posición en la que la contrincante la puede capturar sin que obtengamos nada a cambio de ese sacrificio. Un buen motor debe realizar una buena poda para reducir el tamaño de la lista de movimientos candidatos a ser el mejor. La otra característica es la precisión en la evaluación. Lo que hace único a un motor de ajedrez es la manera, el procedimiento, el algoritmo que tiene para evaluar una posición, es decir, para evaluar una combinación de piezas colocadas en el tablero de una manera dada. Es la función de evaluación.

Una función de evaluación es una fórmula matemática muy complicada. Tiene que tener en cuenta el material que hay en el tablero, las piezas de las que dispone cada bando, es decir, el balance de material. También tiene que evaluar la movilidad de las piezas, si están conectadas entre sí y se apoyan unas a otras. Ver cuántas casillas controlan las piezas propias y las de la contrincante, qué bando tiene el control de las casillas centrales del tablero, que son muy importantes para desplegar estrategias de ataque. Mirar cómo está la estructura de peones, que es a la vez la línea defensiva y de ataque. Ver cuántas islas de peones hay y si estos están aislados, pasados o doblados. Y considerar si el rey está seguro o demasiado expuesto y es vulnerable. Además, todo esto lo tiene que hacer teniendo en cuenta en qué fase se encuentra la partida, si está en la fase de apertura, de juego medio o de final, porque en cada fase lo mejor puede ser distinto.

Elegir cuáles de todas estas características son relevantes en la posición actual, en la fase actual de la partida, no es una tarea sencilla. También es complicado asignar valores numéricos a cada característica y definir la función de evaluación. Todo ello requiere mucho conocimiento sobre el juego.

En el año 2017 Stockfish iba por su versión 8. Tenía un ELO de 3.388. Era un motor de fuerza bruta basado en las técnicas tradicionales de programación. Tenía un modelo de datos eficiente. Su algoritmo de búsqueda era Minimax mejorado con la poda alfa-beta, para poder elegir la mejor jugada sin evaluarlas todas. Su función de evaluación se basaba en heurísticas, que son atajos para sacrificar completitud y exactitud a cambio de velocidad. Se usan porque no es posible, en términos de complejidad, explorar todas las posiciones posibles, por lo

que es necesario hacer uso de una función de aproximación. Con esto se consigue evaluar menos opciones, pero de manera más precisa.

Las heurísticas eran reglas aproximadas que daban a las posiciones valores numéricos basados en el conocimiento humano de personas expertas en ajedrez. Así, se convertían en números los datos más importantes de la posición, ahorrando cálculos y, por tanto, tiempo. La función de evaluación incorporaba el estudio de una gran cantidad de partidas de ajedrez jugadas durante centenares de años. Mucho conocimiento humano disponible gracias al registro de millones de partidas anotadas y accesibles públicamente, condensado en forma de reglas, que reducía la complejidad del algoritmo para alejarse lo más posible del temido NP-completo. Además, para ganar eficiencia, estaba conectado con libros de aperturas y tablas de finales.

Stockfish era un excelente compañero para cualquier ajedrecista que quisiera mejorar su juego. Era un magnífico sistema experto para ayudar en la toma de decisiones. Después de una partida, la jugadora podía usarlo para analizar algunas de las posiciones y comparar la jugada que proponía el motor con la que ella había realizado, buscando los errores. Para mejorar el juego, las ajedrecistas suelen repasar con espíritu crítico los movimientos que han realizado, preguntándose qué hicieron y qué podían haber hecho. En este análisis, que se hace individualmente, con la misma contrincante o con las compañeras del club, es frecuente incorporar un motor. Al fin y al cabo, el ELO del motor es muy superior al humano, así que es de esperar que aporte buenas ideas.

Sacar conclusiones no siempre es fácil. El motor da mucha información numérica que hay que saber interpretar.

Para poder interpretarla, es bueno que el motor juegue lo más parecido posible a un humano. Lo más parecido posible, porque idénticamente es imposible. Y es imposible porque el humano tiene, debe tener, un plan. Un plan puede ser dominar el centro, dominar casillas importantes, atacar el enroque, estropear la estructura de peones al contrario… La jugadora tiene que tener una idea de cómo dominar el medio y el largo plazo de la partida. Al fin y al cabo, para la ajedrecista cada partida es una historia, una narración que mantiene una coherencia interna. En cambio, el motor carece de planes. Juega sin ideas. Solo elige la mejor jugada en cada turno, así que su visión de la partida no es exactamente la de un humano. Pero, para que sirva de ayuda a los humanos, su propuesta, su evaluación, su análisis

debe ser comprensible. Cuando la jugadora pregunta al motor: «En esta posición, ¿qué habrías hecho?», es preciso que ella comprenda la respuesta del motor, que entienda la lógica interna de esa respuesta, porque, si no se entiende, entonces es como una respuesta loca. Puede ser muy acertada, pero la ajedrecista no sabrá cómo incorporarla en sus esquemas mentales, en sus patrones de juego. No le servirá para enriquecer su aprendizaje. Una perfección incomprensible y aburrida no sirve. Un motor de ajedrez no puede salirse por peteneras.

El 5 de diciembre de 2017 se supo que el programa de computador AlphaZero había desafiado a Stockfish 8 a jugar un *match* de 100 partidas. AlphaZero obtuvo 28 victorias, 72 empates (o tablas) y ninguna derrota. AlphaZero era propiedad de la empresa DeepMind, de Google.

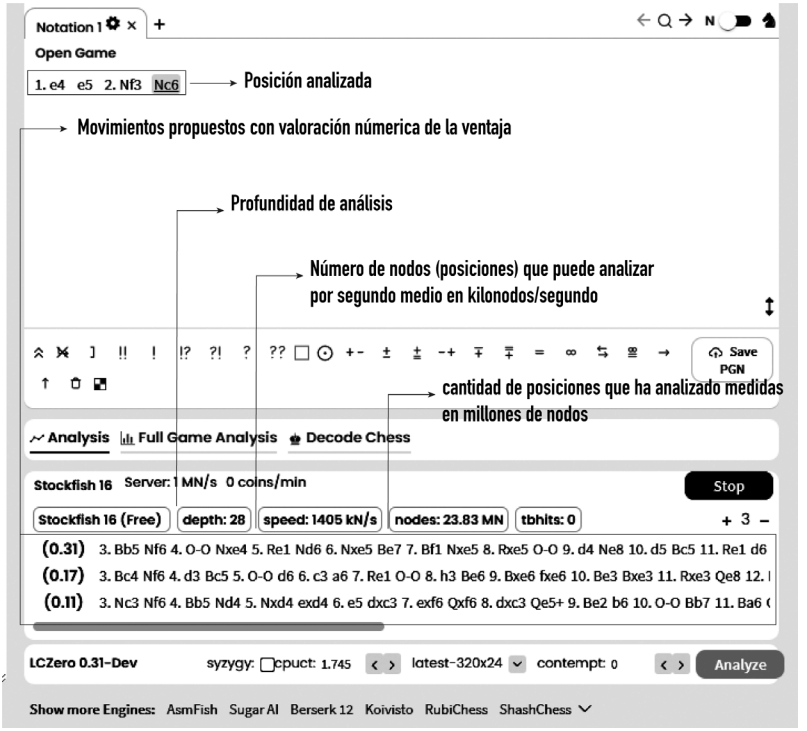


Imagen 7. Información mostrada por un motor de ajedrez tras analizar la posición de la imagen 6 (p. 57) y las diferentes continuaciones posibles. (Fuente: https://official-stockfish.github.io/docs/stockfish-wiki/Download-and-usage.html#download-a-chess-gui).

Impresióname

En el año 2010, DeepMind era una *startup* recién fundada, una de esas empresas emergentes que se centran en alguna novedad tecnológica aún sin explotar y de la que se espera que en poco tiempo alcance un valor de mercado muy alto. En el caso de DeepMind, esa novedad tecnológica era una de las áreas del *machine learning:* el aprendizaje profundo con redes neuronales artificiales organizadas en muchas capas.

La *startup* se financiaba con capital riesgo de inversores como Horizons Ventures y Founders Fund, en cuyas páginas web podemos leer que su propósito es invertir en ciencia y tecnología profundas que puedan impulsar a la humanidad hacia un futuro radicalmente mejor y apoyar el desarrollo tecnológico, ya que la tecnología es el motor fundamental del crecimiento en el mundo industrializado.

El objetivo de DeepMind iba mucho más allá de la creación de máquinas con capacidad de aprendizaje automático. Su ambición era, es, en sus propias palabras, resolver la inteligencia y luego usarla para resolver todo lo demás. Su estrategia es integrar la neurociencia y el *machine learning* bajo el enfoque de las redes neuronales artificiales profundas, y construir máquinas que puedan aprender a resolver problemas para los que no han sido entrenadas específicamente. El objetivo final: una única máquina para miles de problemas distintos. Una caja negra universal para resolverlo todo.

En 2014 Google compró DeepMind por más de quinientos millones de dólares. Google pagó esa cantidad a pesar de que la *startup* no estaba dando beneficios y no tenía ningún producto a explotar. En ese momento, DeepMind ya había desarrollado una máquina capaz de

aprender a jugar de manera autónoma a los videojuegos arcade de la mítica videoconsola de Atari utilizando aprendizaje por refuerzo. El aprendizaje por refuerzo es una técnica de *machine learning* inspirada en la psicología conductista: qué acción escoger para maximizar una recompensa. Bajo el paraguas y la financiación de Google, DeepMind desarrolló AlphaZero.

El 5 de diciembre de 2017 DeepMind publicó un artículo de diecinueve páginas titulado «Mastering Chess and Shogi by Self-Play with a General Reinforcement Learning Algorithm». En él explicaba que su máquina AlphaZero había jugado un *match* de 100 partidas contra el motor de ajedrez Stockfish 8 con el resultado de 28 victorias a su favor, 72 empates (o tablas) y ninguna derrota. Los debates sobre qué es el conocimiento, qué es la inteligencia, qué es aprender y qué es jugar al ajedrez desbordaron los laboratorios de investigación y saltaron a la calle.

Es cierto que las condiciones del *match* eran discutibles. A Stockfish no se le permitió usar bases de datos de aperturas ni tablas de finales, algo que mermaba su potencial. Tampoco se usó su última versión. El tiempo de cálculo para cada movimiento se estableció en un minuto, lo cual penalizaba al motor, que funcionaba por fuerza bruta. Y el hardware sobre el que corría Stockfish no era el óptimo para su desempeño. El artículo de DeepMind no era un *paper* propiamente dicho. No hubo revisión entre pares. Pero, a pesar de todo esto, en el mundo ingenieril de la computación, en los foros de ajedrez y en la calle el resultado del *match* se consideró espectacular. Digno de ser tenido en cuenta.

¿Cómo estaba hecha AlphaZero? Procesaba unas 60.000 posiciones por segundo, muchas menos que los 60 millones que procesaba Stockfish. Su algoritmo de búsqueda del mejor movimiento no era el reconocido Minimax, sino el árbol de búsqueda Monte Carlo. Lo de Monte Carlo viene por referencia al casino, ya que este algoritmo incorpora las probabilidades y la aleatoriedad para encontrar la mejor jugada, aunque precisamente por introducir el azar puede que la solución óptima le pase desapercibida. Monte Carlo se considera un algoritmo más débil que Minimax, ya que, al evaluar posiciones azarosamente, se le puede escapar la buena. Evalúa menos posiciones, pero con más profundidad.

Si todo era un poco peor, ¿entonces cómo ganaba?

La diferencia más significativa entre Stockfish y AlphaZero estaba en la función de evaluación, que es la encargada de calcular cuán buena es una posición (la situación de las piezas de ambos contrincantes en el tablero) respecto a otras posiciones posibles para elegir el camino hacia la mejor de todas.

La función de evaluación de Stockfish estaba programada teniendo en cuenta el conocimiento de la comunidad ajedrecística, así como de una gran cantidad de partidas bien jugadas. Tomaba en consideración todos los factores que un humano considera al evaluar una posición dada. La de AlphaZero tenía una red de neuronas artificiales que, a fuerza de autoentrenamiento, de jugar contra ella misma, había calibrado para calcular la probabilidad estadística de que tal movimiento contribuya a ganar la partida. En su entrenamiento no había utilizado información previa ni partidas jugadas por humanos.

AlphaZero era un experimento. En su camino hacia la resolución de la inteligencia, DeepMind había elegido como terreno de experimentación el de los juegos de turnos alternos entre dos jugadores, como el ajedrez, el go o el shogi. Juegos en los que ambos contrincantes tienen a la vista toda la información relevante en cada momento y en los que no hay aleatoriedad ni incertidumbre. El experimento consistía en construir una máquina, introducir las reglas del juego y el objetivo (en el caso del ajedrez, matar al rey) para que, desde cero, sin conocer nada más sobre el juego, empezara a jugar contra sí misma y fuera aprendiendo sobre la base de jugar su propio juego. Según DeepMind, AlphaZero jugó durante nueve horas 44 millones de partidas contra sí misma, y así ajustó su red neuronal artificial. Máquina autodidacta, aprendió de su propia experiencia, sin que un humano le enseñase nada sobre el juego. Quiero entender que el «Zero», de AlphaZero, alude a ese partir de cero, tabla rasa, obviar toda la experiencia previa, prescindir del conocimiento construido durante cientos de años sobre aperturas, finales, valor de las piezas, estrategias o tácticas. Cero. Hacerse a sí misma desde la nada.

El algoritmo de AlphaZero era genérico. No estaba optimizado para un juego en concreto. Un único algoritmo para aprender a jugar a cualquiera de los juegos. La apuesta de DeepMind era usar el aprendizaje por refuerzo profundo, una de las muchas técnicas de *machine learning*. El aprendizaje por refuerzo profundo se basa en ensayo y error. No requiere interacción humana y funciona bien en situaciones en las que las consecuencias de las acciones se evidenciarán a largo plazo, es decir, cuando hay que ir dando pasos para alcanzar

un objetivo. Para resolver la inteligencia, DeepMind creía que esta era la tecnología adecuada.

Eso sí, el hardware era importante. AlphaZero se había entrenado utilizando un hardware especializado: las unidades de procesamiento por tensores, las TPU. Una TPU es un circuito integrado para acelerar las cargas de trabajo de las máquinas de aprendizaje automático. Estas TPU habían sido diseñadas por Google y eran más de cien veces más rápidas que las GPU de gama más alta. Para entrenarse usó 5.000 TPU. Una vez entrenada, para jugar ya solo requería 4. Entrenar es mucho más costoso que jugar.

Pero, a todo esto, ¿qué pasaba con el ajedrez? Para Google, el *match* era una muestra de fuerza, una exhibición de superioridad para posicionarse frente a sus competidores y tener poder para negociar con los gobiernos la regulación de la inteligencia artificial. El ajedrez era un mero instrumento para afianzar su dominio. Para DeepMind, el ajedrez, un juego con unas pocas reglas muy bien definidas y computacionalmente complejo, era un magnífico laboratorio de experimentación del aprendizaje por refuerzo profundo. ¿Y para las comunidades ajedrecísticas? ¿Para aquellas personas para las que el ajedrez es su vida? ¿Cómo quedó el mundo del ajedrez después de que el rodillo de AlphaZero pasara por encima de Stockfish?

A diferencia de lo que ocurre en otros ámbitos, las personas que practican el ajedrez no quieren ser sustituidas por máquinas. Quieren seguir jugando incluso si juegan mal, ya que, de hecho, todo el mundo juega mal. El motor puede ser un compañero en la toma de decisiones, pero ningún jugador quiere ser sustituido por él, porque jugar es estimulante, es emocionante, proporciona placer.

Como en todo deporte, la práctica del ajedrez es muy variada. No es lo mismo el ajedrez educativo o el terapéutico que el que se juega en las cárceles o en los patios de instituto. No es lo mismo un país que otro. No es lo mismo un torneo para clubs de base que un campeonato mundial. No es lo mismo la afición que el trabajo en los muchos servicios profesionales que se mueven en torno al juego.

Los desarrolladores de DeepMind dijeron que el ajedrez de AlphaZero era alienígena, sobrehumano, de otra dimensión. Desde dentro del ajedrez, lo que sí se hizo fue analizar con esmero en qué consistía esta sobrehumanidad.

El código de AlphaZero no se conocía. La máquina no estaba disponible, DeepMind no la había puesto a disposición pública, no se podía jugar contra ella. Lo único que se tenía eran esas diecinueve páginas del artículo publicado. No se conocían los millones de partidas que jugó contra sí misma para entrenarse. Tampoco las cien que jugó contra Stockfish. En el artículo «Mastering Chess and Shogi by Self-Play with a General Reinforcement Learning Algorithm» se listaban únicamente los movimientos de diez partidas completas.

A pesar de la escasez de información y de que mucho de lo que se cree saber sobre AlphaZero es especulativo, el diagnóstico de todos los entendidos fue unánime: su ajedrez no se parecía en nada al ajedrez humano y, en consecuencia, tampoco al de Stockfish. Era algo esperable, ya que la máquina se había desembarazado de todo el orden conceptual sobre el juego construido por los humanos época tras época. Pero no por esperable dejaba de ser menos impresionante.

Sacrificaba piezas aparentemente a lo loco y hacía movimientos cuya ventaja solo aparecía después de avanzar mucho en la partida. Un juego ininterpretable. Difícil de descifrar para un humano. Un juego no humano, pero ¿no humano significa sobrehumano?

Sobre el consenso de que AlphaZero no era humano se desplegaron un conjunto de valoraciones, de juicios de valor, que naturalizaban una representación del mundo con la que la ideología de la inteligencia artificial conectaba como anillo al dedo. Valoraciones mezcla de tecnología y de ideología.

Se dijo que AlphaZero había necesitado solo unas pocas horas para aprender. Pero no sabemos cuánto esfuerzo, personas y tiempo humanos se emplearon para diseñar su algoritmo. ¿Años?

Se dijo que superaba el conocimiento que a la humanidad le ha costado cientos de años construir. Cuando, en realidad, suponiendo que lo de AlphaZero sea conocimiento, en todo caso será otro conocimiento, ya que su visión del juego se mueve en otro paradigma. Un tipo de juego que va directo a la victoria, algo lógico tratándose de un juego competitivo, pero sin considerar la belleza y el orden en el mientras tanto.

Se dijo que la fuerte dependencia de programas como Stockfish del estilo humano de juego, en lugar de ser una ventaja, era un lastre. Que el conocimiento humano no es más que un prejuicio erróneo e inútil. Pero ¿qué sentido tendría tener motores de ajedrez de los que

no se puede aprender, porque juegan con lógicas que cuesta mucho interpretar? ¿Para qué los querríamos? Lo humano como lastre, como prejuicio, es uno de los hilos argumentales con los que la inteligencia artificial se legitima.

Se dijo que AlphaZero tenía perspicacia e intuición, equiparando a estas facultades el azar y la no exhaustividad de su árbol de búsqueda, así como la ausencia de lógica interna, más allá de la probabilidad estadística. Intuición de AlphaZero mejor que razonamiento de Stockfish.

Se dijo que AlphaZero comenzaba con un lienzo en blanco y con capacidad de aprendizaje, igual que el bebé al nacer. Como si los bebés humanos aprendieran de la nada, por sus propios medios, sin madres, sin el esfuerzo y los cuidados de la comunidad que les ofrece tramas de significados en los que los conocimientos toman sentido.

Se dijo que su juego era innovador y creativo, que no era lineal, que usaba estrategias no inmediatamente evidentes para los humanos o para los motores de ajedrez. La innovación como valiosa en sí misma es otro de los hilos argumentales para su legitimación. Pero ¿para qué sirven innovaciones que no comprendemos, si las capacidades humanas no pueden sacar provecho de ellas?

Se dijo que no era una calculadora Turing robusta que avanzaba a través de miles de millones de pasos, sino una máquina verdaderamente pensante, con percepciones similares a las del ingenio humano. Argumento discutible, porque ese supuesto ingenio similar al humano en realidad desarrolla un tipo de juego muy divergente del humano.

Se dijo que la máquina juega sin afecciones, que no siente la presión, que no le afecta el estrés de la competición. Lógico, igual que le ocurre a Stockfish.

Y a la vista de todo este debate, ¿cuál fue la conclusión?

Bueno, es cierto que AlphaZero ganó más ELO que Stockfish, pero el ELO de Stockfish ya era mucho más alto que el del campeón mundial y muchísimo más alto que el de la mayoría de los jugadores medios. ¿Para qué queremos una máquina con un ELO mucho mayor? ¿Por qué deberíamos enterrar un motor que se distribuye como software libre, que se programa colaborativamente y que puede correr en cualquier ordenador personal para endiosar una máquina privativa que necesita un hardware especial, que no

está disponible y que, si lo estuviera, no serviría para mejorar masivamente el juego humano?

Pero, por otra parte, ¿por qué deberíamos rechazar una tecnología que puede mejorar, superar, las que estamos usando?

El 9 de enero de 2018 se anunció el nacimiento de Leela Chess Zero (Lc0), un motor de ajedrez programado colaborativamente, gratuito y distribuido como software libre. Surgió como iniciativa de algunos de los desarrolladores de Stockfish y su objetivo es verificar los métodos de AlphaZero. Lc0 utiliza una red neuronal entrenada para jugar al ajedrez. En su página web las desarrolladoras dicen que las redes actuales de Lc0 han superado la fuerza de AlphaZero.

En octubre de 2020, Stockfish anunció que su versión 12 incorporaba en su función de evaluación la red neuronal Efficiently Updatable Neural Network (NNUE), que puede activarse o desactivarse para apoyar el análisis tradicional.

La comunidad ajedrecística, que está organizada y que ama el ajedrez, va a lo técnico, determina su hoja de ruta, marca su propia hoja de ruta.

Y, más allá de Google, en los laboratorios de *machine learning*, ¿cómo quedó la cosa?

El experimento AphaZero mostró que se podían usar redes neuronales para abordar problemas que antes se consideraban demasiado complejos como para ser resueltos con un enfoque computacional.

Mostró que un diseño de propósito general adaptable y versátil (no especializado en ajedrez) permitía a la máquina aprender a jugar a ajedrez tan bien o mejor que un diseño específico para ese juego.

Que usando técnicas parecidas a las que se usan para el reconocimiento de imágenes la máquina podía aprender a identificar la similitud entre situaciones, en el tablero, aunque estas fueran distintas en cuanto a la posición concreta de las piezas. Algo así como aprender la intuición.

Que hay maneras de alcanzar objetivos que discurren al margen de la lógica humana establecida. Una especie de creatividad artificial.

Y se especula con que después del aprendizaje AlphaZero hubiese elaborado, manejase, construcciones abstractas como las diagonales abiertas, la estructura de peones, la seguridad del rey…, como si tuviese capacidad de abstracción.

En la actualidad, DeepMind ha metido la cabeza en el mercado de la salud y ha desarrollado escáneres para la retina que pueden detectar el glaucoma, la retinopatía diabética o la degeneración macular asociada a la edad. Y su proyecto AlphaFold, para desarrollar fármacos y hacer biología sintética para crear nuevas proteínas especialmente diseñadas para aplicaciones industriales o medioambientales, está muy avanzado.

Desembróllame

Si el combate entre AlphaZero y Stockfish escenifica la lucha entre lo nuevo y lo viejo, ¿eso significa que hay dos bandos? Incluso, yendo más lejos, ¿eso significa que hay una inteligencia artificial buena y otra mala?

La génesis de la inteligencia artificial se puede situar en los años cincuenta, cuando disciplinas como la computación, la teoría de la información y las ciencias cognitivas convergieron en una pregunta: ¿se puede crear inteligencia fuera del sistema humano cerebro-mente?

Para las ciencias de la información la pregunta abría un apasionante mundo de experimentación sobre los límites. Para las ciencias cognitivas suponía tener un laboratorio de pruebas en el que realizar experimentos maquínicos que ayudarían a comprender mejor qué es la inteligencia humana. Construir máquinas que piensen y comprender qué es pensar. Unas ciencias aprendiendo con otras.

Enseguida surgieron las dos inspiraciones que hemos señalado: la logicista (también llamada simbólica o cognitivista) y la conexionista. Ambas compartían la premisa de que el sistema cerebro-mente responde a un modelo computacional y, por tanto, es susceptible de ser replicado por computadoras. Ambas eran materialistas. Se proponían estudiar la mente desde presupuestos puramente físicos. Pero ahí terminaba lo compartido. Más allá de este punto, las dos inspiraciones divergían.

La inspiración logicista ponía el punto de mira en el funcionamiento racional del sistema cerebro-mente. Por decirlo de una manera simple, estaba fascinada con el cerebro izquierdo. Si en la

computadora mente-cerebro la mente es el programa de ordenador, el software, y el cerebro es la quincalla, el hardware, el logicismo se quedaba con el programa de ordenador.

Entendía que la inteligencia humana viene a ser una expresión del razonamiento lógico. Razonar es procesar la información. Esa información, para que pueda ser procesada, tiene que estar bien organizada. No caben los cajones de sastre. La información tiene que estar estructurada y las reglas tienen que estar establecidas. Para organizar la información se necesitan símbolos, representaciones. Hay que poder representar los objetos, las ideas, las reglas, las relaciones. Y eso se hace manejando símbolos. De ahí que a esta escuela se la conozca también como simbolista.

Una manera metafórica de aproximarnos al manejo de símbolos es observando el juego simbólico de los niños. La niña juega a coches con coches de juguete, pero también puede jugar a coches con cajas de cartón o con pinzas de la ropa. La caja de cartón o la pinza de la ropa simbolizan el coche. No hace falta tener un coche de verdad para jugar a coches. Tanto da lo que tengas en la mano: si quieres jugar a coches, basta con ver en ello un coche y hacerlo funcionar como un coche. Algo parecido ocurre dentro de un programa de ordenador que quiere replicar el razonamiento humano sobre algún dominio siguiendo la propuesta simbolista. Dentro de ese programa de ordenador, tiene que haber símbolos de las entidades que está manejando. Desde este enfoque, disponer de lenguajes de programación específicos que manejasen bien la lógica era muy importante.

El simbolismo adoptaba un enfoque analítico, de arriba abajo. La hipótesis es que, si se resuelve lo complejo, quedará resuelto lo más simple. La inteligencia funciona mediante deducciones. Va de lo general a lo particular. Para que la máquina pueda emular un razonamiento, se le tienen que haber introducido explícitamente todos los conocimientos humanos disponibles relativos al ámbito en cuestión. Una inteligencia artificial deductiva, a lo Descartes.

Por su parte, la inspiración conexionista ponía el punto de mira en el funcionamiento biológico del sistema mente-cerebro. El quid de la cuestión está en el cerebro, no en la mente. Lo que hay que comprender no es la capacidad de razonamiento, sino la capacidad de aprendizaje. Y la capacidad de aprendizaje se gesta en la quincalla, en el hardware. De la computadora mente-cerebro, el conexionismo

desechaba la mente y se quedaba con el cerebro. Por eso se dice que el conexionismo es de inspiración biológica.

¿Qué hay en el cerebro? Muchas neuronas y muchas sinapsis. Pues bien, diseñemos y construyamos neuronas artificiales que se puedan conectar entre sí. Hagamos redes neuronales. La unidad básica de una red neuronal artificial es la neurona artificial. Hay que enseñar a esas neuronas a que, dependiendo de la información que les entre, o bien desencadenen una reacción en cascada en otras neuronas o bien se desactiven, se inhiban. Hay que enseñarles a tener un comportamiento. Hay que entrenar la red. La hipótesis era que pequeñas unidades actuando de modo autónomo, descentralizado e inteligente harían emerger inteligencias muy complejas.

Aprender significa modificar las conexiones, reforzando o debilitando las sinapsis. En las redes neuronales artificiales, cada sinapsis tiene un peso que determina su capacidad de influenciar a otras neuronas. Una neurona suelta puede muy poco, pero la potencia de la red consiste en el procesamiento en paralelo: todas las neuronas actuando a la vez. Ninguna neurona es responsable del resultado final, pues ese resultado es fruto de la actividad conjunta de todas las neuronas actuando de consuno. Ni que decir tiene que todo esto eran desarrollos físicos y matemáticos.

Aunque el conexionismo se inspiraba en el cerebro entero, tenía predilección por el lado derecho, el creativo. Adoptaba un enfoque de síntesis, de abajo a arriba. Vamos a ir de lo particular a lo general. La hipótesis es que, aprendiendo de muchos casos concretos, particulares, después la red podrá generalizar. El aprendizaje funciona mediante inducciones. Para que la máquina pueda emular un aprendizaje, hay que darle casos, no leyes. A partir de un montón de casos, ya irá aprendiendo cuáles son las leyes generales. Una inteligencia artificial inductiva, a lo Hume.

Al principio ambas inspiraciones avanzaban en paralelo. ¿Qué puede haber de malo en desarrollar dos visiones, cuando no se sabe de antemano cuáles son las potencias y los límites de cada una? Al fin y al cabo, habían surgido entre colegas que compartían los mismos ambientes tecnológicos y académicos. Tenían discrepancias, sí, como las hay en todas partes, pero no había necesidad de hacer sangre. Sin embargo, el pacto de no agresión no duró mucho, porque enseguida surgió un conflicto competitivo: había que luchar por conseguir financiación, mucha financiación.

Para conseguir financiación, hay que convencer a los financiadores. Y eso supone no solo defender que lo propio es lo bueno y lo mejor, sino también que lo ajeno es lo malo y lo peor. Hay que desplegar una crítica al adversario, desmantelar sus argumentos, evidenciar sus puntos débiles y, en último término, hundir su credibilidad.

En los años cincuenta, el paradigma dominante era el logicista. Defendía nociones que en esa época estaban en boga: estructurar, controlar, dirigir de arriba abajo, organizar según jerarquías… Manejaba valores asentados. Esta concepción racionalista de la inteligencia era criticada por el conexionismo: el conocimiento, en su emerger, es intuitivo y empírico, no racional. La propuesta conexionista era disruptiva y desafiante. Enfatizaba la naturaleza cooperativa y paralela de los procesos cerebrales. Construía redes no jerárquicas, sin centro, asumiendo que pudieran llegar a tener una dinámica impredecible. La información no estaba concentrada en un único lugar. Se despedazaba, se fragmentaba, y los fragmentos se repartían entre todas las neuronas. Así, por distribución y cooperación, la complejidad se haría asumible. Lo que hacían los logicistas podía funcionar, pero no había que llamarlo inteligencia.

En los años sesenta ambas propuestas habían alcanzado resultados y generaban expectativas. El logicismo trabajaba en la construcción de sistemas expertos tipo Eliza, que se considera el primer programa de ordenador capaz de procesar el lenguaje natural. Eliza se comunicaba con las usuarias mediante texto (teclado y pantalla) y proporcionaba una conversación terapéutica. Era una psicóloga virtual, que ahora llamaríamos chatbot. Demostraba la posibilidad de que una máquina interactuara con humanos procesando el lenguaje natural. No cabía duda de que era un gran hito. Inteligencias artificiales generales, no especializadas en nada concreto, como el ordenador HAL de *2001: Una odisea del espacio*, podrían estar listas en unos veinte años.

Por su parte, el conexionismo podía poner encima de la mesa el Mark I Perceptron, como hemos visto. Ese autómata eléctrico se basaba en el uso del perceptrón, que es un modelo de red neuronal con el que se podían reconocer patrones. Primero aprendía y después era capaz de generalizar clasificando información nueva que no había visto nunca, que en este caso concreto eran fotografías de rostros humanos. Otro gran hito que hacía pensar que máquinas que pudieran caminar, hablar, ver, escribir, reproducirse y tener consciencia estarían, a no muchos años vista, al alcance de la tecnología. Máquinas

capaces de percibir, reconocer e identificar su entorno sin necesidad de control humano.

El perceptrón estaba financiado por la Oficina de Investigación Naval de Estados Unidos. Pero para continuar la investigación hacía falta más dinero. Hacía falta el dinero de la Agencia de Proyectos de Investigación Avanzada (ARPA). Había que convencer a ARPA de que la investigación en redes neuronales artificiales era prometedora.

Fue entonces, en 1969, cuando logicistas muy reconocidos publicaron *Perceptrons: An Introduction to Computational Geometry*. El libro era un golpe a la línea de flotación de las redes neuronales artificiales. Los autores demostraban matemáticamente que el perceptrón tenía grandes limitaciones, como así es. Con un rigor matemático irreprochable, rezumaba la valoración de que era un sinsentido invertir en la investigación de algo que aspiraba a ser una inteligencia artificial pero ni siquiera podía distinguir si una figura era conexa o inconexa y no era capaz de programar algo tan simple como una función lógica OR exclusiva (XOR).

Es cierto que no podía. Para lograrlo, había que desarrollar una arquitectura en capas, muchas capas de neuronas, y había que tener un algoritmo de retropropagación *(backpropagation)* para que las neuronas artificiales ajustaran mejor sus pesos. No eran, en realidad, límites absolutos. Se trataba solo de dar más tiempo y más recursos. Todo eso está en las redes neuronales actuales y podría haberse llegado a ello quizás en no mucho tiempo si desde el logicismo no se hubiera sembrado con tanta beligerancia una sospecha escéptica sobre esta línea de investigación.

La ARPA decidió explícitamente apoyar la inteligencia artificial simbólica y, no muy entrados los años setenta, las propuestas conexionistas se quedaron en el sueño de construir máquinas capaces de aprender.

Entonces ¿en los años setenta la inteligencia artificial buena era la conexionista y la mala la simbólica, y ahora es al revés: la mala es la conexionista y la buena es la simbólica?

Las tecnologías vienen envueltas en ideas. Confrontan ideas. En los años sesenta, la concepción de un enjambre compuesto por pequeñas unidades autónomas, autoorganizado, colaborativo y capaz de hacer cosas sin una figura central de mando era una crítica a la supuesta necesidad de un autoridad jerárquica sabelotodo. El conexionismo

desafiaba al logicismo. ¿Quién ha dicho que los expertos son trigo limpio *per se*? El racionalismo es autoritario y también tiene sesgos, muchos sesgos. Una red neuronal sin director de orquesta era una idea potente y liberadora. Recordemos que el Mayo del 68 fue en 1968.

Hoy el capitalismo ha puesto a trabajar para sí versiones descafeinadas de esta idea. El neoliberalismo propugna la desregulación total. Por supuesto, no le gustan los enjambres autoorganizados y colaborativos, pero sí se sirve de una multitud de *riders* que van de acá para allá a golpe de lo que un descorporeizado algoritmo autoritario demanda o, más bien, exige. Desestructurar, desorganizar para explotar más y gobernar mejor. Es en este contexto en el que trato de interpretar la contienda entre AlphaZero y Stockfish. Un contexto en el que la existencia y la supervivencia de comunidades no es baladí.

El *machine learning* es una ingeniería envuelta en el celofán de la inteligencia artificial; la inteligencia artificial opera como ideología, y en la misma medida que como tecnología. La ideología sirve para ofuscar las tecnologías subyacentes y para naturalizar y legitimar ideas, ideas que colonizan el pensamiento y hacen que cuando suena el término «inteligencia artificial» vengan a la cabeza prodigios portentosos como ChatGPT o AlphaZero, y en cambio nadie piense en Stockfish o en cualquiera de los motores de ajedrez usados por los jugadores como sistemas expertos.

La ideología de la inteligencia artificial se lleva el agua a su molino. El *match* escenificó la lucha de lo nuevo contra lo viejo. Lo viejo, Stockfish. Lo nuevo, AlphaZero. Tecnologías en confrontación, pero sobre todo luchas por delinear el imaginario popular sobre la inteligencia artificial, por establecer el parteaguas que hace que AlphaZero caiga del lado de la inteligencia artificial y Stockfish no. Por identificar la inteligencia artificial con lo alienígena y lo sobrehumano.

AlphaZero vino para decirnos lo bueno que es quitarte de encima el lastre de la memoria, olvidar todo lo anterior; lo bien que va a funcionar una escuela sin maestras, donde el alumnado aprende por sí mismo generando sus propios itinerarios; lo obsoleto que es el esfuerzo, la fuerza bruta, porque la genialidad te llegará por sí sola si te reinventas y eliges un buen estilo de vida; que la falta de estructuras, es decir, el imperio de estructuras desestructuradas, cuyas leyes internas no son explícitas, es bueno para la creatividad; que hay que zafarse de las restricciones para adaptarse a las reglas del juego, de los juegos cambiantes, de cualquier juego; que hay que ir por la vida

sin afecciones, que hay que arriesgar; que puedes desprenderte de algunas piezas, de todo lo que te estorbe, sin problema si con ello vas directo a tu objetivo, a vencer…

Entonces ¿es que las tecnologías dan lo mismo, porque al final todo es pura ideología? Por supuesto que no. Las tecnologías son importantes. Y las ideas también. Comprender las tecnologías significa poderles quitar el celofán, la envoltura legitimadora. Mirarlas cara a cara y poder envolverlas, si es preciso, en un nuevo envoltorio.

El mítico artículo «La tiranía de la falta de estructuras», publicado en 2003 y dirigido a colectivos sociales, plantea en clave interna el debate sobre la recuperación de técnicas tradicionales que son útiles, aunque no perfectas. Algo parecido podría ocurrir con la *Neuro-symbolic AI*, una propuesta de inteligencia artificial híbrida que combina redes neuronales artificiales con sistemas lógicos en un intento de integrar lo mejor de cada mundo.

Para mantener líneas de investigación abiertas, hace falta ganar espacio en el terreno de las ideas. Y, por supuesto, hace falta financiación. No es cierto que para que haya evolución tecnológica tenga que haber guerras. Como se vio en la pandemia de COVID-19, basta con que haya voluntad; es decir, recursos.

Propaganda

El 20 de marzo del 2023, el Future of Life Institute publicó una carta abierta titulada *Pause Giant AI Experiments*. La carta hacía un llamamiento a «todos los laboratorios de inteligencia artificial para detener inmediatamente durante al menos seis meses el entrenamiento de sistemas de IA (inteligencia artificial) más potentes que GPT-4». El 14 de marzo, seis días antes, OpenAI había lanzado GPT-4. La carta la firmaron miles de empresarios, investigadores, empleados de grandes corporaciones y otras personas influyentes.

El Future of Life Institute es una institución sin ánimo de lucro cuya misión es dirigir la tecnología transformadora hacia el beneficio de la vida y alejarla de los riesgos extremos a gran escala.

¿Cuáles son esos riesgos a gran escala? Siempre según Future of Life, respecto a la inteligencia artificial «lo peor está por llegar. Las empresas buscan activamente la inteligencia artificial general que pueda realizar muchas tareas igual o mejor que los humanos. Prometen que esto traerá beneficios sin precedentes, como curar el cáncer o acabar con la pobreza mundial. Pero más de la mitad de los expertos en IA creen que hay una posibilidad entre diez de que esta tecnología provoque nuestra extinción». ¡Una posibilidad entre diez de provocar la extinción de la humanidad! No es moco de pavo. Y siguen explicando que este riesgo no es particular, sino general.

A largo plazo, no deberíamos fijarnos en un método concreto de hacer daño, porque el riesgo procede de la propia inteligencia avanzada. Pensemos en cómo los humanos dominan animales menos inteligentes sin depender de un arma concreta o cómo un programa

de ajedrez de inteligencia artificial derrota a jugadores humanos sin depender de una jugada específica.

Actualmente no tenemos forma de saber cómo actuarán los sistemas de IA, porque nadie, ni siquiera sus creadores, entiende cómo funcionan. La seguridad de la IA se ha convertido en una preocupación generalizada. Los expertos y el público en general están unidos en su alarma ante los riesgos emergentes y la necesidad apremiante de gestionarlos.

Puede sonar bonito, pero algo no cuadra si la institución cuya misión es alertar de que las inteligencias artificiales pueden extinguir a la humanidad está constituida por las mismas personas que las están construyendo y uno de sus asesores es precisamente Elon Musk.

Elon Musk es un magnate multimillonario que tiene un patrimonio de más de doscientos cincuenta mil millones de dólares. Entre otras, fundó Neuralink, una empresa de neurotecnología que desarrolla interfaces cerebro-computadora implantables en el cerebro humano que buscan —según dice él mismo— una simbiosis total con la inteligencia artificial. Fundó OpenAI, cuyo proyecto más destacado es ChatGPT, el popular chat basado en un modelo de lenguaje por inteligencia artificial, aunque se desvinculó de esta empresa en 2019. Fundó SpaceX, para viajar al espacio y colonizar Marte. Fundó The Boring Company, para hacer túneles bajo tierra y que el tráfico de las ciudades sea subterráneo. Entró en Tesla, cuya misión era revolucionar la industria automotriz produciendo coches eléctricos eficientes de alta gama y que ahora fabrica coches autónomos para que funcionen con asistentes a la conducción y pilotos automáticos sin necesidad de conductor, usando inteligencia artificial a saco. Fundó xAI con el objetivo —dice— de comprender la verdadera naturaleza del universo y, entretanto, entrar en el negocio de la IA generativa con Grok, un chat para competir con el Bard de Google, el Copilot de Microsoft y, por supuesto, el ChatGPT de OpenAI. También fundó X Corp, la sucesora de la red social Twitter, que ahora ofrece Grok a sus usuarios *premium*.

¿Por qué un magnate que está en el negocio de las inteligencias artificiales, como Elon Musk, promueve y difunde manifiestos apocalípticos en contra de ellas?

Las ingenierías de la inteligencia artificial se están acelerando para desarrollar productos rentables. Los gigantes tecnológicos de la industria están financiando batallas competitivas por colocar productos, especialmente inteligencias artificiales generativas del tipo

de ChatGPT, en el mercado de masas. Detrás del consumo de masas hay otras industrias no tan visibles, pero no por ello menos jugosas, como la de producción de chips con procesadores especializados por parte de empresas de hardware como Nvidia o Intel, o la de servicios de alojamiento en la nube para que las inteligencias artificiales estén siempre disponibles, lo cual es negocio para empresas como Microsoft o Amazon.

En esta batalla competitiva por dominar el mercado de masas, tan importante es llegar el primero para colocar el producto como controlar el celofán que lo envuelve.

En febrero de 2023, la agencia de seguridad del tráfico de Estados Unidos instó a Tesla para que tomara medidas urgentes para corregir los errores de su sistema de asistencia a la conducción *Full Self Driving*, una inteligencia artificial que hace que un automóvil pueda conducir, acelerar, frenar y cambiar de carril de manera automática, sin conductor. El sistema había causado ya al menos 273 accidentes, cinco de ellos con consecuencias mortales. Tesla anunció que retiraría del mercado más de 362.000 vehículos equipados con *Full Self Driving*. Un golpe al producto, pero no pasa nada si no se daña el celofán. El producto remonta si el celofán queda intacto.

¿Qué es el celofán? El celofán es la aureola con la que se envuelve el producto. No se trata solo del empaquetado tecnológico, sino de la legitimidad que lo envuelve. Cómo se presenta, cómo lo percibe la gente. El celofán es la emoción positiva, la carga de positividad que acarrea el producto, su necesidad, su bondad. El envoltorio es lo mejor del regalo. Hay que construir la validez, la licitud del producto, impresionar, cautivar, hacerlo deseable. Hay que borrar del mapa las preguntas que los sistemas sanitarios se plantearon en torno al MYCIN. No más preguntas. «Nosotros, los magnates, controlaremos las preguntas. Ya os diremos lo que se puede preguntar y lo que no. Nosotros, los magnates, controlaremos vuestras ideas y opiniones. Ya os diremos lo que es una inteligencia artificial y lo que no. Y, por supuesto, ya os decimos de entrada que inteligencia artificial es solo lo que hacemos nosotros».

Y ¿cómo se hace para controlar el celofán? Pues lanzando mensajes apocalípticos abstractos que ocupen el espacio que podrían tomar los debates situados.

En julio de 2017, en la reunión de verano de la Asociación Nacional de Gobernadores (NGA) de Estados Unidos (celebrada en Rhode

Island), Elon Musk declaró que la inteligencia artificial es el mayor riesgo que enfrentamos como civilización y que hasta que la gente no vea a los robots matar a personas por la calle no se entenderán los peligros de la inteligencia artificial. La cuestión es que, mientras la prensa y los tertulianos hablan sobre la hipotética posibilidad de que en un futuro haya robots por la calle matando gente, dejan de hablar de los coches de Tesla equipados con la inteligencia artificial asistente a la conducción *Full Self Driving*, esos que sí van ya por la calle provocando accidentes y, en ocasiones, muertes. Mientras nos asustan imaginando una futura inteligencia artificial que nos matará, dejamos de pensar en cómo se están produciendo ahora violaciones de la privacidad, sesgos, precariedad laboral, impactos medioambientales, posverdades y distorsiones deliberadas de la realidad o privatizaciones del conocimiento. Ya lo dicen los del Future of Life Institute: no nos fijemos en los métodos concretos de hacer daño. Cada método concreto viene envuelto en celofán y ese celofán hay que mantenerlo intacto.

Es una especie de diálogo tal que así:

— Oye, que la IA puede exterminar a la humanidad.

— ¡¿Qué dices?! Eso es imposible, no va a ocurrir.

— Ah, vale, pues entonces no hay problema. ¡Que entre a saco!

La carta del Future of Life Institute o las declaraciones de Elon Musk no son la primera ni la última profecía apocalíptica.

Informáticos influyentes como Geoff Hinton —un investigador de algoritmos de aprendizaje profundo reconocido y galardonado que ha trabajado en Google hasta los setenta y cinco años— alertan del peligro real de que la IA resulte ser un desastre. Hinton dice que las máquinas van a ser mucho más inteligentes de lo que pensaba. Y que tiene miedo. Cuando en 2015 se le preguntó por qué continúa con las investigaciones a pesar de sentir tan hondas y graves preocupaciones sobre estas tecnologías, respondió: «Te podría dar los argumentos habituales. Pero la verdad es que la perspectiva de descubrir es demasiado dulce». Demasiado dulce.

El 4 de junio de 2024 trece trabajadoras y extrabajadores de OpenAI, Anthropic y DeepMind publicaron una carta abierta en la que afirmaban que los riesgos que plantea la inteligencia artificial «van desde un mayor afianzamiento de las desigualdades existentes, pasando por la manipulación y la desinformación, hasta la pérdida de control de los sistemas autónomos de IA, lo que podría provocar la extinción

humana». En la carta defendían la necesidad de regulación y hacían un llamamiento a las empresas de élite para que no represalien a sus empleados cuando estos plantean públicamente sus inquietudes relacionadas con los riesgos. «Mientras no haya una supervisión efectiva de estas grandes empresas por parte de los gobiernos, los antiguos y actuales empleados se encuentran entre las pocas personas que pueden obligarlas a rendir cuentas ante la ciudadanía. Sin embargo, los amplios acuerdos de confidencialidad nos impiden expresar nuestras preocupaciones, salvo a las mismas empresas, que puede que no estén por la labor de abordar estas cuestiones. Las medidas habituales de protección del denunciante son insuficientes, porque se centran en la actividad ilegal, cuando buena parte de los riesgos que nos preocupan no están todavía regulados. Algunos de nosotros tenemos miedos fundados a diferentes formas de represalia, dado el historial de casos así en todo el sector».

Este llamamiento se produce en un contexto de caza de talentos por parte de las empresas que desarrollan inteligencia artificial, para no quedar rezagadas en la carrera. En el mundo académico, los doctorados se vacían porque los estudiantes reciben ofertas de trabajo en la IA que ponen por delante mucho dinero y muchos recursos. La oportunidad de trabajar en investigaciones y desarrollos que están cambiando el mundo es una tentación demasiado dulce.

Detrás de estas preocupaciones aterradoras puestas en boca de esos ingenieros, que son todo menos inocentes, puede resonar la propia inestabilidad que crea entre las grandes corporaciones del sector un crecimiento tan acelerado. La carta abierta de Future of Life se publicó unos pocos días después de que OpenAI lanzara GPT-4. Pedía detener durante al menos seis meses el entrenamiento de sistemas de IA potentes. Suena a: «Oigan, paren un poco todo esto y dennos tiempo a las demás empresas a reorganizarnos, que nos estamos quedando atrás».

Además de la inestabilidad, están los riesgos en la seguridad, en la estabilidad financiera de los mercados. Como dice Elon Musk, las máquinas podrían comenzar una guerra publicando noticias falsas, robando cuentas de correo electrónico y enviando notas de prensa falsas, solo con manipular información. Ahora ya no habla del exterminio de la humanidad, sino del riesgo de un potencial colapso de la información con consecuencias reales para los mercados financieros y la economía, donde la difusión de rumores, verdaderos o falsos, puede desplomar las cotizaciones en bolsa. Los capitalistas temen que las

inteligencias artificiales generativas generen escenarios volátiles y, de alguna manera, demandan algún tipo de regulación, frenar un poco, no explotarlas a lo loco, donde cada cual entiende por «a lo loco» lo que vaya en contra de sus intereses. Un lío.

Pero, mientras esos mismos empresarios, investigadores o ingenieros que están creando y explotando inteligencias artificiales toman el megáfono para alertarnos de los riesgos existenciales que estamos asumiendo como humanidad, se deja de poner atención en las situaciones concretas de discriminación algorítmica, de perpetuación de patrones de discriminación racial, social o de género, de exacerbación de estereotipos, de desinformación, de costes medioambientales o de precariedad laboral.

La estrategia de las élites del sector es enunciar los riesgos como algo abstracto y al mismo tiempo preservar el celofán de los productos concretos. Hablar de riesgos de exterminio de la humanidad omitiendo el entramado de relaciones de poder en medio del cual la inteligencia artificial se está implantando. Levantar gigantes que oculten los molinos.

La consecuencia es grave. Las mismas élites controlan el desarrollo y las críticas. De esta manera ofuscan completamente las posibilidades no ya de una soberanía tecnológica popular, sino de una mínima comprensión real de qué son las inteligencias artificiales, cuáles son las tecnologías base, quién las está desarrollando, cómo se están implantando, quién se está beneficiando…

Los pronósticos apocalípticos asientan la opinión de que inteligencia artificial es aquello que puede tomar decisiones autónomas contra la humanidad. Confusión. En este imaginario, el pobre Stockfish queda fuera, así como también quedarían fuera MYCIN, el Mark I Perceptron, las máquinas con bucles de realimentación y tantas y tantas otras máquinas que técnica y socialmente son inteligencias artificiales, pero que no están en la cresta de la ola.

Con esta confusión, sin comprensión de lo que es, resulta imposible construir un pensamiento crítico autónomo.

Qué es

«Inteligencia artificial» no es un término tecnocientífico. Es un término de *branding*. Es impactante. Vende muy bien, pero describe poco. Así que no hay una única definición de inteligencia artificial aceptada universalmente.

Para la reciente Ley de Inteligencia Artificial de la Unión Europea, que es el reglamento regulador, la Comisión Europea proponía como definición: «Software que se desarrolla empleando una o varias de las técnicas y estrategias que figuran en el anexo I y que puede, para un conjunto determinado de objetivos definidos por seres humanos, generar información de salida como contenidos, predicciones, recomendaciones o decisiones que influyan en los entornos con los que interactúa».

El anexo I establecía las siguientes técnicas y estrategias:

a) Estrategias de aprendizaje automático *(machine learning),* incluidos el aprendizaje supervisado, el no supervisado y el realizado por refuerzo, que emplean una amplia variedad de métodos, entre ellos el aprendizaje profundo *(deep learning).*

b) Estrategias basadas en la lógica y el conocimiento, especialmente la representación del conocimiento, la programación (lógica) inductiva, las bases de conocimiento, los motores de inferencia y deducción, los sistemas expertos y de razonamiento (simbólico).

c) Estrategias estadísticas, estimación bayesiana, métodos de búsqueda y optimización.

En las negociaciones que rodearon la elaboración de esta ley, al final el Parlamento Europeo hizo cambios y la cosa ha quedado en que un sistema de inteligencia artificial es un sistema basado en máquinas que está diseñado para funcionar con distintos niveles de autonomía y que puede mostrar capacidad de adaptación tras su despliegue, y que, para objetivos explícitos o implícitos, infiere, a partir de la información que recibe, cómo generar resultados tales como predicciones, contenidos, recomendaciones o decisiones que pueden influir en entornos físicos o virtuales.

En un texto legal que va a regular la implantación y el uso de la inteligencia artificial en Europa, cada palabra se mira con lupa y es el resultado de negociaciones tensas entre el sistema político y la industria, en las que las entidades defensoras de la justicia algorítmica también intentan hacer oír su voz.

Observemos que ninguna de las definiciones anteriores se acerca a la definición más extendida, que dice que la inteligencia artificial es una rama de la informática que desarrolla programas capaces de emular procesos propios de la inteligencia humana. La expresión «inteligencia humana» no aparece por ninguna parte. No sirve para afinar un texto regulatorio. No es útil para delimitar por ley el contorno de lo que entra y de lo que queda fuera.

Por contra, los cambios entre la primera definición y la segunda vienen a decir que la inteligencia artificial, a efectos de regulación, no se puede definir por las técnicas con las que se programa (aprendizaje automático, programación lógica, estadística, etcétera). Además, la autonomía es un rasgo importante: se quita la frase «objetivos definidos por seres humanos» y se añade que puede funcionar con distintos niveles de autonomía para objetivos implícitos o explícitos, es decir, que el sistema en cierto modo puede ser autónomo para fijar sus objetivos, que no tienen que ser siempre los definidos por seres humanos. Y un tercer cambio es que entre los entornos con los que interactúa se incluyen los físicos y los virtuales, como por ejemplo los entornos de realidad virtual o mundos virtuales tipo metaverso.

Llevando el texto a un lenguaje coloquial y quitando todos los «puede», queda que una inteligencia artificial es un sistema con un cierto nivel de autonomía (funciona sin intervención humana), que tiene una cierta plasticidad para adaptarse a entornos cambiantes, que recibe información, que la maneja para generar resultados que responden a objetivos (predicciones, contenidos, recomendaciones,

decisiones…) y que esos resultados influyen en el exterior de ese sistema. Y esto independientemente de cómo se haya construido o programado.

Bien. Y, en este saco, ¿qué es lo que entra? Montones de cosas. Tomemos, por ejemplo, un antivirus. Funciona sin intervención humana (una vez instalado), se adapta a entornos cambiantes (reconoce nuevos virus), recibe información (el programa que está escaneando), genera un resultado (permitir o no la instalación de ese programa) que responde a un objetivo (mantener el ordenador libre de virus) y el resultado influye en el exterior (el ordenador donde está instalado). Luego un antivirus es una inteligencia artificial. ¿Sorprendente? Si resulta sorprendente, es porque todo esto se mueve en un batiburrillo sin elaborar, mezcla de marketing, sensacionalismo, mensajes apocalípticos, futurología y ciencia ficción, que impide hacerse con una comprensión crítica y situada de lo que realmente supone este salto tecnológico que, sin duda, es abstracto y complejo.

En realidad, si nos atenemos a la definición de que una inteligencia artificial es un programa de ordenador capaz de emular procesos propios de la inteligencia humana, el antivirus también entra de lleno. Distinguir si un código de programa de ordenador es malicioso o no es una tarea que haría una persona programadora estudiando el código, y para eso se necesita inteligencia. El antivirus emula esa inteligencia detectando patrones. Es una inteligencia artificial detectora de patrones. Así que la definición en sí no afecta al grueso de lo que se entiende por una inteligencia artificial y, si la Unión Europea no se basa en las tecnologías utilizadas, es para no pillarse los dedos respecto a tecnologías que se pudieran aplicar en el futuro y que por no estar en la lista jugarían al escapismo.

El saco de las inteligencias artificiales es muy grande y las hay de todo tipo. Las hay *fashion*, como Siri o Alexa, y las hay que no tienen ningún *glamour*, como el reconocimiento de matrículas de vehículos en las zonas urbanas de altas emisiones o el detector de *spam* en el correo electrónico. Unas están hechas por grandes corporaciones, como AlphaZero, y otras están hechas por comunidades de programadores, como Stockfish. Hay inteligencias en servidores en Internet, como ChatGPT, y otras que se instalan y configuran en el ordenador personal y se usan sin necesidad de conexión a Internet, como Jan (jan.ai). Jan es una inteligencia artificial de tipo chatbot —alternativa a ChatGPT— que se usa en local, en el ordenador personal, por lo que está enfocada en la privacidad y no cede ningún dato personal. Además, se

distribuye como software libre. Unas se desarrollan en superordenadores de corporaciones y otras en superordenadores públicos como el Marenostrum, que está en el Centro Nacional de Supercomputación, en Barcelona. Unas son públicas, como las del Centro Europeo de Previsiones Meteorológicas a Medio Plazo, que desarrolla métodos numéricos para predicciones meteorológicas de medio plazo y las ofrece a los Estados miembros, y otras son privadas, como el servicio meteorológico GraphCast, propiedad de DeepMind.

Unas tienen mucho impacto en la generación de desigualdades y de injusticia social, como las de reconocimiento de rostros de personas en los controles fronterizos, y otras se celebran como un gran avance, como las que detectan pecas cancerígenas en la piel. Ambas son sistemas de reconocimiento de imagen, pero sus consecuencias palmarias no tienen nada que ver. Pueden ser técnicamente, por dentro, muy parecidas y socialmente, por fuera, algo completamente distinto.

Entonces ¿lo técnico no es importante? Al fin y al cabo la Unión Europea ha evitado definir las inteligencias artificiales en función de las tecnologías empleadas para desarrollarlas.

Lo técnico es importante, aunque no es lo único importante. La Unión Europea no quiere referirse a tecnologías concretas para que la ley, que tanto ha costado promulgar, no quede obsoleta al día siguiente, o para que la industria no se escude en recovecos muy técnicos para eludir su aplicación. Pero lo técnico es importante.

A consecuencia del salto tecnológico que se produjo con el control de la energía eléctrica, ahora las casas están llenas de enchufes donde conectar aparatos. La utilidad de un aparato será la misma, pero es importante saber si la energía que consume es de una fuente fósil o de una fuente renovable, ¿no? El origen de la energía no cambia el uso, pero socialmente es una pregunta muy relevante. Para cooperativas de consumo como Som Energia, esa es la pregunta esencial.

Cuando la conexión a Internet pasó de ADSL a fibra óptica, este cambio no fue solo un mero aumento de velocidad. Conllevó un cambio muy importante en la gobernanza de la infraestructura. Mientras que la ADSL utilizaba el par de cobre de las líneas telefónicas y todas las operadoras ofrecían servicio sobre una única estructura de hilos de cobre, una estructura única común, cuando llegó la fibra óptica cada operadora tiró su propio cable. El subsuelo está cuajado de cables que se despliegan en paralelo, uno al lado del otro, cada uno de una operadora distinta. Al cambiar la tecnología de conexión a

Internet cambió el modelo de gobernanza de la infraestructura. Para guifi.net, un proyecto tecnosocial cuyo objetivo es la creación de una red de telecomunicaciones abierta, libre y neutral, basada en el modelo de comunes, la pregunta sobre la gobernanza de la infraestructura es una pregunta esencial.

Y así podríamos seguir. No es lo mismo que el agua del grifo sea de pantano que de una desaladora. No es lo mismo que el agua de riego sea subterránea que de trasvase. Riega igual, pero no es lo mismo. Para cada caso habrá que valorar, habrá que ver. Para levantar un pensamiento crítico, lo técnico de cada cosa concreta no es un detalle menospreciable a delegar en los especialistas. El pensamiento crítico debe abrirse al conocimiento técnico concreto, superar el desconocimiento para quitar la magia y poder decidir cómo queremos relacionarnos con las tecnologías, con cada una de las tecnologías concretas tal y como cada una se está estableciendo. Un pensamiento crítico situado que esté a la altura de las transformaciones dejará de oscilar entre la fascinación y la amenaza, y recuperará protagonismo social. Para ello, la pregunta sobre cómo funciona cada cosa concreta es esencial.

Es esencial y no es sencilla, porque los enfoques tecnológicos y las condiciones sociales son dinámicos y van cambiando y, por tanto, no es posible tener respondidas las preguntas de una vez por todas.

Vértigo

Carta a Toni

Querido Toni, ¿cómo estáis?

Cuando acabe de escribir el libro seguiremos hablando de filosofía, ¿verdad? Los libros se acaban, pero los problemas no ;-)

No veas lo que estoy disfrutando con el libro que me prestaste, ese de Kuhn, el de la revolución copernicana. Nunca había leído los textos de esos astrónomos empeñados en explicarse cómo funciona lo que hay más allá de la Tierra. Y la verdad es que cuando los leo no entiendo ni jota. Pero el libro explica muy bien lo del cambio de paradigma y lo de que en un determinado momento histórico unas ideas son concebibles mientras que otras no lo son. Los antiguos ya conocían muy bien las ventajas de colocar el Sol en el centro del sistema en lugar de colocar la Tierra. Todo quedaba más simple y encajaba mejor. Pero eso, sencillamente, no podía ser. Era ¡inconcebible!

También explica muy bien el hilo que va desde Copérnico hasta Newton. ¡Nada menos que trescientos años! No veas el tiempo que necesitaron las mentes pensantes para terminar de comprender y aceptar cómo es el sistema solar.

Pero lo que más me flipa es que, de esos filósofos científicos que fueron proponiendo objeciones a la idea preconcebida de que la Tierra era el centro del universo, ninguno de ellos tenía la intención de subvertir nada ni de rebelarse contra la autoridad de nadie, ni de poner en cuestión la cosmología heredada. De hecho Galileo dedicó su obra al papa de Roma y Newton creía en un Dios creador eterno, infinito, absolutamente perfecto. Sin embargo, pusieron patas arriba la relación entre Dios y su creación, rompieron la unidad entre ciencia y filosofía y desbrozaron el camino del método experimental. La tecnología se empezó a desarrollar sin contenciones, hasta el infinito.

Según el libro, la revolución copernicana termina cuando Newton afirma y demuestra que la gravitación es una fuerza universal. Es muy impresionantes eso de ¡universal!

Significa que las leyes de nuestro mundo terrenal y las leyes del cosmos son las mismas. No hay nada sublime en el universo que le dé mayor perfección, que lo eleve por encima de la precariedad terrenal, de aquí donde todo se degrada, se desorganiza y muere. Las leyes físicas son las mismas para todo quisqui. Aquí en la Tierra y más allá en cualquier rincón de universo. Mismas leyes. ¡Toma ya!

Con esto de la inteligencia artificial me pregunto si cuando nos resistimos a creer que una inteligencia artificial sea inteligente estaremos haciendo como los que no querían quitar la Tierra del centro del universo. Pobres mortales que somos… No sé si nos estamos negando a cambiar la posición hegemónica de lo humano. No sé si estaremos diciendo que no puede ser porque es ¡inconcebible! y ya está.

Los de la cibernética empezaron a decir que máquinas y humanos se podían explicar con los mismos modelos. No hay nada sublime en lo humano que le dé mayor perfección, que lo eleve por encima de la artefactualidad sin espíritu. Las leyes son las mismas, dijeron. ¡Da miedo!, ¿no?

Bueno, a la próxima lo comentamos. Veme preparando más libros, je, je.

Besos.

3. Estrategias

Cómo son

Como hemos visto, con las tecnologías actuales hay dos estrategias principales para desarrollar inteligencias artificiales: el enfoque simbólico y el enfoque conexionista. Aunque se está investigando en una inteligencia artificial neurosimbólica (Neuro-symbolic AI), que podría aunar lo mejor de ambas y superar las limitaciones de cada una de ellas, esta tercera vía todavía no está lo suficientemente madura, por lo que en la práctica solo hay dos estrategias.

El enfoque simbólico se basa en reglas escritas por personas expertas y no necesita grandes cantidades de datos. Las reglas están bien establecidas y representan el conocimiento humano estructurado sobre determinada materia. Son claras y están predefinidas. Es una estrategia idónea para diseñar sistemas expertos.

Por ejemplo, para programar una inteligencia artificial que calcule el precio del alquiler mensual de una vivienda las reglas podrían ser:

Regla 1: Si tiene más de cincuenta años de antigüedad, entonces el alquiler baja 300 €.

Regla 2: Si está en el centro de la ciudad y tiene piscina, entonces el alquiler se incrementa en 500 €.

Regla 3: Si tiene menos de 50 m², entonces el alquiler se calcula a 16 €/m².

Regla 4: El alquiler no puede ser menor de 800 € ni mayor de 3.000 €.

Etcétera.

El enfoque simbólico es ordenado, representa un razonamiento comprensible para los humanos. De hecho, encapsula ese razonamiento y eso tiene ventajas: destripa la caja negra. Como responde a razonamientos lógicos, el sistema puede explicar el porqué de sus decisiones. Su comportamiento es verificable y explicable. Todo lo contrario de una caja negra, que ofusca lo que hay en su interior, oculta los detalles internos e imposibilita acceder a su funcionamiento.

Esta ventaja, la comprensibilidad y explicabilidad, es muy valorada en escenarios de riesgo alto, en los que se necesita confianza en el proceso, tener la seguridad de que funciona. MYCIN era simbólico.

El otro enfoque es el conexionista. El enfoque conexionista se basa en modelos estadísticos o en el uso de redes neuronales artificiales. Utiliza grandes cantidades de datos y no utiliza reglas. A partir de un entrenamiento que consiste en procesar montones de datos, «simplemente» se espera que las relaciones emerjan. Se utiliza para el reconocimiento de imágenes, sistemas de recomendación, chats conversacionales, etcétera. El Mark I Perceptron era conexionista.

Aunque en principio estas dos estrategias son aplicables, en la actualidad el enfoque conexionista es el que se está llevando el gato al agua. Ofrece muy buenos resultados y además, para una gran variedad de objetivos, es el más barato, pues requiere menos tiempo de trabajo humano de alta remuneración (personas expertas, programadoras, etcétera). En la historia y en la academia se reconoce que la inteligencia artificial tiene una rama simbólica, pero para la industria y para la cultura popular a día de hoy esta rama se considera anticuada.

Veamos ahora cómo es una inteligencia artificial conexionista por dentro.

Supongamos que una profesora calcula la nota media del curso haciendo una media ponderada en la que el resultado de cada examen trimestral puntúa el 20 % y el trabajo final puntúa el 40 %.

El modelo matemático de la nota media sería parecido a lo que muestra la imagen 8, donde:

- x_1, x_2 y x_3 son los resultados de los tres exámenes trimestrales
- x_4 es el resultado del trabajo final
- w_1, w_2 y w_3 valen 0,2
- w_4 *vale 0,4*
- y la salida (n) es la nota redondeada, sin decimales.

Si una alumna saca un 6 en el primer examen, un 4 en el segundo, un 2 en el tercero y un 8 en el trabajo final, entonces:

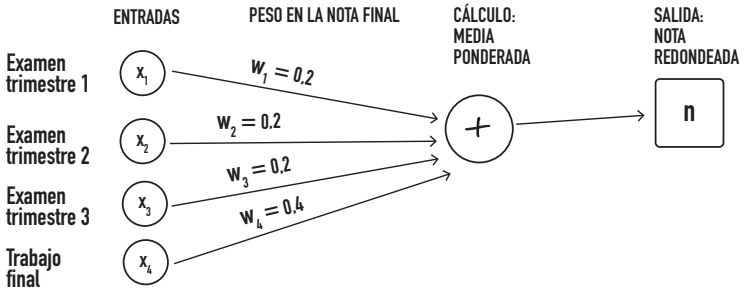


Imagen 8. Modelo matemático de la nota final.

$x_1 = 6$
$x_2 = 4$
$x_3 = 2$
$x_4 = 8$

El modelo matemático calcula:

$(x_1 * w_1) + (x_2 * w_2) + (x_3 * w_3) + (x_4 * w_4)$

$(6*0,2) + (4*0,2) + (2*0,2) + (8*0,4) = 5,6$

Al resultado 5,6 se le aplica el redondeo y la salida sería 6.

Este modelo matemático es un perceptrón. Su esencia es la misma que la del Mark I Perceptron, aunque con una diferencia muy grande: el Mark I aprendía.

¿En qué consistiría el aprendizaje en este modelo matemático que calcula la nota final como una media ponderada redondeada?

Sería darle como entrada datos con las notas de distintas personas, decirle cuál es la salida correcta y que él mismo aprenda, calcule cuáles son los valores para w_1, w_2, w_3 y w_4.

Por ejemplo, entrenarlo con este conjunto de datos:

x_1	x_2	x_3	x_4	Salida
6	4	2	8	6
6	4	2	6	5
5	8	2	4	5
3	4	5	6	5
6	8	2	7	6
…	…	…	…	…

Y esperar a que afine los valores que ponderan cada nota parcial, es decir, cuál es su peso (w_1, w_2, w_3 y w_4), sin darle de antemano las reglas de ponderación.

¿Y qué sería esperar que las relaciones emerjan? En este caso concreto nada, porque no hay relaciones. Cada nota parcial tiene un peso independientemente de los valores de las otras notas parciales. No hay relaciones entre ellas. Una nota parcial influye, pesa en la nota final según su propio valor independientemente de los valores de las otras.

Pero imaginemos un caso más complejo: una profesora puntúa a su alumnado de manera intuitiva y global. Puntúa los exámenes, pero luego pone la nota que le parece justa. ¿Qué tiene en cuenta para poner una nota justa? Tiene en cuenta la situación que hay en casa, el nivel cultural de la familia, qué notas tuvo en cursos anteriores, cómo es su grupo de amistades en el aula, la calidad de su alimentación, su salud, su estado emocional… La profesora, tomando en consideración de modo intuitivo todos estos factores en su conjunto y en sus interrelaciones, y teniendo en cuenta las notas parciales, pone una nota final a ojo.

Supongamos que queremos hacer una inteligencia artificial que aprenda el modo de evaluación de la profesora y que lo haga según la técnica del aprendizaje automático supervisado. Estamos asumiendo que ese modelo es matematizable, lo cual no es una asunción menor. Asumiendo esto, que implica llevar todo a números, en primer lugar habría que hacer una lista de todo lo que hay que tener en cuenta como entrada para el entrenamiento del modelo (lo que tiene en cuenta la profesora): situación en casa, nivel cultural de la familia, notas anteriores, etcétera. Habría que hacer una lista de los factores que ella contempla de manera intuitiva. ¿Cuál sería el tamaño de la lista, es decir, el tamaño de los datos de una entrada? Uhm, pues… para este ejemplo pongamos que 26, más las 4 notas parciales: un total de 30 factores para la entrada (cada alumna). Después de tener esa lista de factores, habría que preparar un conjunto de datos de entrenamiento pidiéndole a la profesora que tome alumnas concretas, en situaciones concretas, y ponga número, valores numéricos, a cada uno de los factores de entrada, que tome también sus notas parciales y la nota final que puso. La profesora, tirando de memoria, construiría una tabla como la que se muestra en la siguiente página, en la que las filas son las notas, los factores y la nota final, y las columnas representan personas.

El conjunto de datos de entrenamiento podría ser una tabla de este estilo. Pero tenemos un problema.

Alum.	Alum. 1	Alum. 2	Alum. 3	…
Nota primer examen x_1	6	7	5	…
Nota segundo examen x_2	8	5	6	…
Nota tercer examen x_3	4	5	5	…
Nota trabajo final x_4	6	5	4	…
Situación en casa x_5	8	4	9	…
Nivel cultural de la familia x_6	9	4	3	…
Notas de cursos anteriores x_7	4	4	6	…
…	…	…	…	…
Nota final n	5	6	6	…

Como todo es un poco intuitivo, otra profesora puede cuantificar los factores de forma diferente. Y una tercera profesora lo mismo. No se puede entrenar una inteligencia artificial para que procese datos de entrada que cada cual pone a su aire ni se puede esperar que las intuiciones de todas las profesoras converjan armónicamente.

Entonces se podría adoptar otro enfoque. Olvidemos los factores. Nos quedamos solo con las notas y, en lugar de factores, vamos a tomar documentos que contengan informes: el informe de los servicios sociales, el de la pediatra, el de la tutora… Entrenaremos la inteligencia artificial con las notas y los informes. Los informes ya contienen información sobre los factores, solo que en lugar de ser información rígida y numérica es flexible, descriptiva y textual. Pero para eso está la inteligencia artificial. Ella hará que las relaciones emerjan.

Así que se entrenaría la red neuronal artificial dándole, alumna por alumna, los datos de entrada (las notas parciales y los informes disponibles) y la salida que corresponde a esas entradas (la nota final). La red neuronal calcularía los pesos, las *w*.

Sin embargo, no es tan sencillo, porque un esquema como el que muestra la imagen 9 (p. 98) no puede funcionar, ya que no se puede equiparar una nota parcial con un informe.

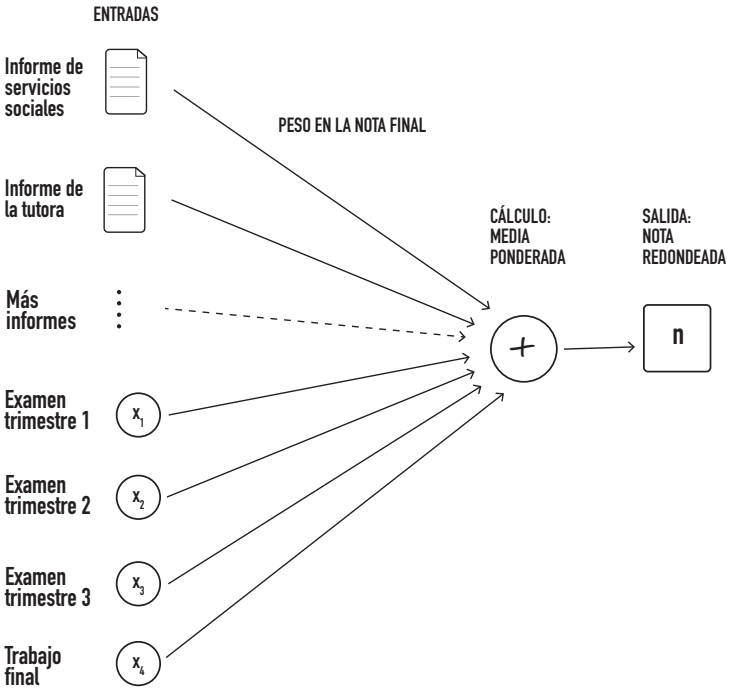


Imagen 9. A ojos vista, este modelo no puede funcionar, porque no se puede equiparar una nota parcial con un informe.

Pero, además, un modelo tan simple no sirve cuando de lo que se trata es de captar las relaciones.

La profesora que hace una valoración global intuitiva sí que ve relaciones entre los factores. Precisamente por eso, la profesora capta los datos globalmente, intuyendo las relaciones que hay entre ellos y no por separado uno a uno. Por ejemplo, una familia puede tener un nivel cultural bajo, pero, si en casa hay muy buen ambiente y la familia valora mucho la cultura, hay buenas condiciones para el aprendizaje. Pero ¿qué pasa si esa niña no está bien alimentada? Una mala alimentación produce cansancio y no facilita la concentración. ¿Cómo interactúa este factor con el resto? En una situación con un mínimo de complejidad, los factores se compensan, se solapan, se amplifican o se equilibran unos con otros, según relaciones difíciles de modelizar matemáticamente. Es por eso que la profesora no calcula una fórmula, sino que capta los datos de forma intuitiva, y precisamente por eso tendría sentido hacer

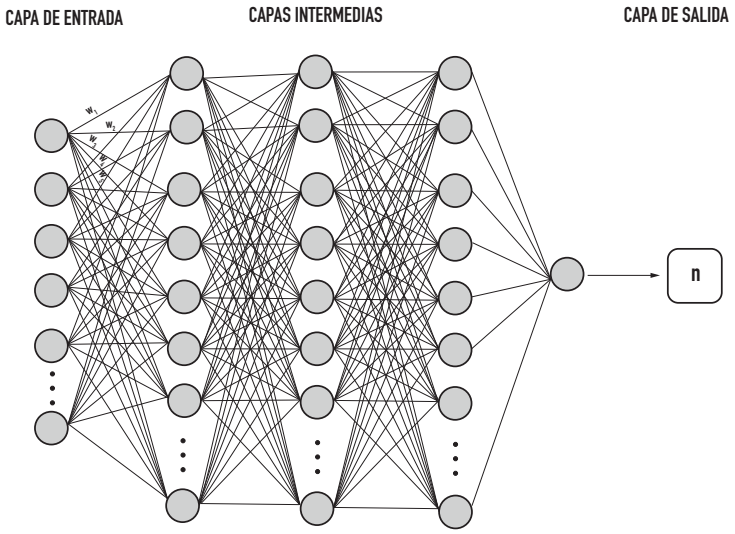


Imagen 10. En lugar de un único y simple perceptrón, muchas capas con muchas neuronas cada capa.

una máquina de inteligencia artificial conexionista en lugar de calcular la nota con una fórmula matemática fija. Dejemos que la máquina haga emerger las relaciones.

Para que la máquina pueda hacer que emerjan las relaciones, hay que usar el modelo (la red) más complejo. Eso se hace añadiendo capas intermedias de neuronas artificiales. En lugar de un único y simple perceptrón, muchas capas con muchas neuronas cada capa (imagen 10).

Y a esa red con muchas capas y muchas neuronas le pondremos delante un procesador de informes capaz de leer el lenguaje natural y estimar los valores numéricos de los factores a partir del texto. Le pondremos delante otra inteligencia artificial entrenada para tal fin. La primera capa, la capa de entrada, tendrá 30 neuronas: 26 para las características más 4 para las notas parciales.

Las capas intermedias pueden tener, y de hecho es lo habitual, un número de neuronas mucho mayor. Supongamos que nuestra red tiene tres capas intermedias. En la imagen 11 (p. 100) se muestran las tres capas, pero no se ven todas las neuronas, que están indicadas por los puntitos. Supongamos entonces que en cada capa hay 500 neuronas.

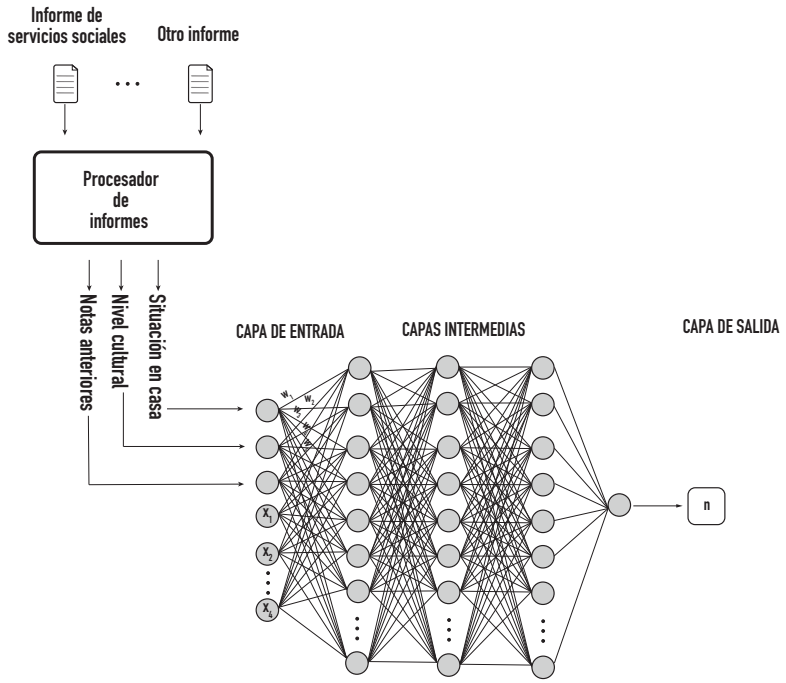


Imagen 11. Red neuronal con un procesador de informes que estima los valores numéricos de los factores a partir de los textos.

Cada una de las líneas de la imagen de la red que va de bola a bola debe tener su peso, su *w*. En este ejemplo, se tendrían que definir 30*500*500*500 valores para las distintas *w*. Eso da un total de 3.750 millones de w. Esta inteligencia artificial tendrá 3.750 millones de pesos, es decir, de parámetros. El entrenamiento de la red consiste en que ella misma calcule el valor que tiene que tener cada uno de los 3.750 millones de parámetros. Este entramado de conexiones tan denso, organizado en capas, es lo que permite a la máquina hacer que las relaciones emerjan.

El modelo de ChatGPT-3 tiene 175.000 millones de parámetros y fue entrenado con cientos de miles de millones de palabras.

Y cuando la máquina esté entrenada, testada y validada, si se mira dentro, ¿qué se verá? Lo que se verá es una matriz de 3.750 millones de números, algo parecido a la imagen 12 (p. 101).

¿Y qué significa cada uno de los números de la matriz? En realidad, nada. Cada número, cada parámetro, cada peso, no tiene ningún sentido por sí mismo. Los números operan en conjunto, todos a la vez, y en conjunto hacen emerger las relaciones que hay entre los datos de entrada, que se plasman en la salida, en la nota final.

$$\begin{bmatrix} 1 & -1 & 0{,}5 & 0{,}2 & 5 & -3 & 3 & 1 & 4 & \dots \\ 0{,}8 & 1 & -1 & 4 & 2{,}5 & 3 & 3{,}1 & -2{,}5 & 3 & \dots \\ 5{,}5 & 4 & 2 & -6{,}7 & -8 & 6 & 0 & 12 & 1 & \dots \\ -3 & 5 & 2{,}1 & 8 & -4{,}2 & 1 & 6 & 7 & -2 & \dots \\ 1 & -1 & 0{,}5 & 0{,}2 & 5 & -3 & 3 & 1 & 4 & \dots \\ 0{,}8 & 1 & -1 & 4 & 2{,}5 & 3 & 3{,}1 & -2{,}5 & 3 & \dots \\ -3 & 5 & 2{,}1 & 8 & -4{,}2 & 1 & 6 & 7 & -2 & \dots \\ 5{,}5 & 4 & 2 & -6{,}7 & -8 & 6 & 0 & 12 & 1 & \dots \\ 1 & -1 & 0{,}5 & 0{,}2 & 5 & -3 & 3 & 1 & 4 & \dots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots \end{bmatrix}$$

Imagen 12. Parámetros de la red neuronal.

En el enfoque conexionista, la inteligencia artificial ajusta los parámetros del modelo (las w) a partir del conjunto de datos con el que ha sido entrenada. Esto implica que los datos de entrada deben seleccionarse con mucho cuidado. Los datos contienen la inteligencia, las leyes del modelo, pero de una manera tan intrincada que esas leyes, en realidad, no las vamos a conocer. La máquina no podrá explicar cuáles son esas leyes ni cómo las está aplicando. Gestiona grandes cantidades de datos difíciles de entender para los humanos, pero sus cálculos son ciegos. Maneja bien la incertidumbre, pero no es comprensible. Funciona, pero no es explicable.

Caja negra.

Enfoques

Para hacer una inteligencia artificial, elegir entre el enfoque simbólico o el conexionista depende del problema a resolver, de los medios disponibles y del contexto cultural, científico, ingenieril y empresarial.

El enfoque simbólico se fundamenta en la lógica formal. Utiliza lenguajes formales para representar el conocimiento y algoritmos para manejar esas representaciones. Hace inferencias y deducciones sobre la base de reglas predefinidas. Representa un razonamiento comprensible para los humanos o, mejor dicho, los humanos representan en el sistema experto sus propios razonamientos. De un modo u otro, el sistema imita el pensamiento, los estados mentales.

Asume que la inteligencia es razonamiento. Los problemas se pueden resolver recorriendo una especie de laberinto lógico en el que, si se sigue el camino correcto, se halla la solución. La gracia está en identificar cuál es ese camino. Una vez resueltos los problemas matemáticos de la explosión combinatoria, es decir, de la gran cantidad inmanejable de opciones posibles, es un enfoque muy bueno para crear sistemas expertos estrechos, en los que el dominio —el área de especialización relacionada con el entorno de la aplicación— está muy acotado. Ciertamente, supone el abandono de la idea de que es posible construir inteligencias artificiales genéricas que resuelvan los problemas de distintos dominios.

Su gran ventaja es que es explicable y una de sus desventajas es que hay que dar mantenimiento a su base de conocimientos, pues el conocimiento sobre un tema determinado varía continuamente. Hay que mantenerlo de por vida, y la industria no está muy interesada en este enfoque de inteligencia artificial a la antigua usanza, que

requiere compromisos a largo plazo. Sin embargo, sigue siendo un enfoque viable en determinados contextos en los que comunidades humanas estables cristalizan su propio conocimiento o en los que se puede garantizar una actualización continua.

Por su parte, el enfoque conexionista se basa en el aprendizaje automático. Para el aprendizaje automático se pueden utilizar distintos tipos de algoritmos, como por ejemplo los modelos estadísticos, los árboles de decisión o las redes neuronales. Entonces, si hay muchos enfoques distintos, ¿por qué se habla tanto de las redes neuronales? La razón es que ahora mismo son lo que más se utiliza. Las redes neuronales lo están petando.

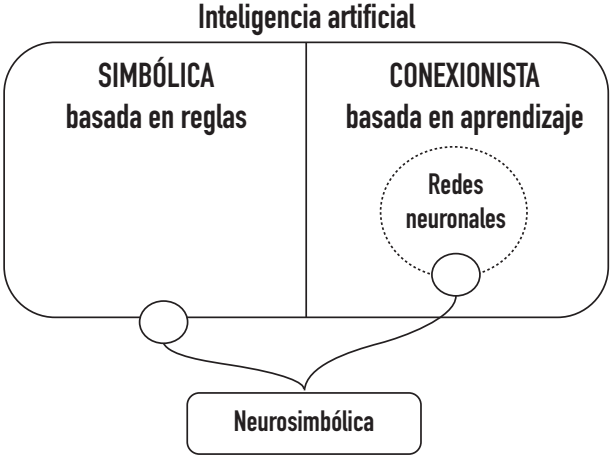


Imagen 13. Clasificación de los enfoques para hacer una inteligencia artificial.

Las redes de neuronas artificiales hacen muchos cálculos y los hacen muy rápido, sin representar simbólicamente el problema que tratan de resolver. Son modelos de inspiración biológica. Modelos bioinspirados en el cerebro humano, pero también en otras expresiones de inteligencia colectiva, de inteligencia de enjambre, como la de las colonias de hormigas, las bandadas de pájaros o los bancos de peces, donde cada individuo del enjambre realiza una tarea muy pequeña al tiempo que transmite y recibe información de una parte del resto, y la resultante es una tarea grupal que se ha llevado a cabo

sin el mandato de un líder organizador. No necesitan director de orquesta.

Detrás de este enfoque late un cierto retorno a la cibernética, a los sistemas que se autorregulan por sí mismos, aunque se elimina la intencionalidad. En lugar de buscar la representación del conocimiento humano, se busca la interconexión de muchos elementos individuales no inteligentes (neuronas artificiales) que, a fuerza de reforzar o debilitar conexiones, convierten esas conexiones en aprendizaje. Se asume que la inteligencia surge por asociación, por las conexiones entre elementos interdependientes simples no jerárquicos, organizados en red, a partir de una actividad distribuida y en paralelo. Más que reproducir el pensamiento humano, lo que intentan es aumentarlo. Llegar a donde este no llega.

Una vez resueltos los problemas matemáticos de la retropropagación con la formulación de algoritmos que mueven la información, dentro de la red neuronal, adelante y atrás para que esta aprenda de sus propios errores y ajuste bien el cálculo de los pesos de cada conexión (las w), en un vertiginoso paroxismo de ensayo y error, son muy buenas para detectar patrones en estructuras de datos muy complejas, como, por ejemplo, detectar, en un TAC, un cáncer que el ojo humano no atina a ver.

Una de sus ventajas, en algunos tipos de aprendizaje automático, es que tienen cierta flexibilidad para adaptarse a entornos cambiantes, autoajustándose a partir de sus propios aciertos y errores.

Las inteligencias artificiales conexionistas en general tienen menor tasa de error que las simbólicas. Aciertan más. Además, son más baratas de producir, porque requieren menos esfuerzo de programación: tienen menos código. Es cierto que necesitan más datos, pero obtener datos puede no ser caro. Con las relaciones laborales actuales y con las técnicas de recolección de datos, muchos datos y poco código es más barato que mucho código y pocos datos.

También es cierto que para entrenarse necesitan más infraestructura de hardware, más maquinaria física, pero esta cantidad de maquinaria física es asumible en infinidad de pequeños y medianos proyectos que tienen objetivos muy acotados. Solo se dispara cuando se busca crear inteligencias artificiales megalómanas tipo ChatGPT, pero la mayoría de proyectos de inteligencia artificial no son *chats gpts*.

El *boom* de la inteligencia artificial basada en redes neuronales se está produciendo porque, comparada con los sistemas anteriores, da mejores resultados y es más barata. Sin embargo, tiene dos grandes problemas. El primero es que necesita grandes conjuntos de datos para ser entrenada y que el contenido de esos datos es crítico para determinar lo que va a aprender; los datos pueden ser baratos, pero tienen que ser buenos. Y el segundo, todavía más crítico, es que no son explicables. Nadie, ni siquiera quienes las han programado, saben explicar en qué se basan para tomar sus decisiones. Son opacas. Por ejemplo, en el caso de la profesora, la inteligencia artificial relacionará los factores, pero no podremos aprender cómo lo está haciendo.

La falta de explicabilidad plantea problemas éticos. ¿Es aceptable utilizar una inteligencia artificial conexionista para prescribir medicación a los pacientes sin poder justificar sus decisiones?

A pesar de su falta de explicabilidad, el enfoque conexionista es ahora mismo el más explotado, hasta el punto de que en sociedad se presenta —amplificado por las operaciones propagandísticas de las grandes corporaciones— como el único posible. En su relato, que no en su tecnología, se reviste con los valores de lo nuevo, entendiendo por lo nuevo una mezcla entre la potencia autoorganizativa de las redes distribuidas y las corrientes neoliberales de desregulación.

Las redes neuronales artificiales vienen envueltas en un celofán disruptivo: representan la liberación de las ataduras formales y el fin del sometimiento a las figuras de autoridad de la tradición. Defendiendo su inspiración biológica, sostienen que la máquina aprende a partir de experiencias, o de su observación del entorno, o de su interacción con él. En definitiva, de su exposición al mundo.

Aun aceptando la metáfora de las experiencias (no sé si se puede afirmar que una máquina tiene experiencias), este enfoque omite, no considera, no le parece relevante todo lo que el humano aprende de modo simbólico. El niño, la niña, aprende a hablar en la lengua que la rodea a fuerza de observar e interactuar. Escuchando y diciendo frases concretas, aprende su lengua materna y ya puede generalizar creando palabras y comprendiendo frases que no ha oído nunca. Pero de mayor, si quiere aprender otra lengua, puede ser útil un enfoque mixto basado en la práctica de la lectura y la conversación, pero también en el aprendizaje de reglas; por ejemplo, las reglas de acentuación.

No aprendemos que el fuego quema por habernos quemado unas cuantas veces. Sabemos que el fuego quema incluso si nunca nos hemos quemado. No lo hemos aprendido por experiencia, sino por transmisión simbólica a través del conocimiento de nuestros mayores (que seguramente tampoco se quemaron nunca), expresado mediante el lenguaje. Igual que no podemos aprender a leer y escribir ni a hacer cálculos por experiencias, sino manipulando símbolos.

En el *match* entre Stockfish y AlphaZero, el primero representa a un jugador veterano, estudioso, meticuloso y especializado que ha dedicado su vida, con esfuerzo y tesón, a adquirir conocimientos sobre el ajedrez, partida a partida, análisis a análisis. AlphaZero representa a una comunidad (de neuronas) joven que, desde cero, se autoorganiza para aprender sin necesidad de jerarquías ni autoridades expertas. Aprende de sí y por sí mismo, sin libros, sin pizarras. Stockfish representa los valores de la disciplina y la constancia. Aprende por transmisión de conocimientos y por acumulación incremental, aceptando la autoridad de lo anterior. El segundo representa los valores de las redes de cooperación sin mando. Aprende por curiosidad y por su propia creatividad, a fuerza de trastear.

Esta película es una de las narrativas posibles. Ensalza el poder de las redes y de la autoorganización sin mando, pero elude la inexplicabilidad, obvia la opacidad. ¡Ojo con las narrativas! Proyectan utopías, valores que van más allá de las tecnologías en sí. Son metáforas construidas socialmente. Son lecturas. Son solo ideas, pero ideas influyentes.

En los años sesenta las tecnologías, en esencia, eran las mismas que las actuales. Sin embargo, las narrativas eran otras muy distintas. El conexionismo, que por esa época fue denostado y se dejó de financiar, es ahora el enfoque dominante. Su narrativa conecta muy bien con la idea de que por fin es posible, y es bueno, una academia sin maestros, sin contenidos estructurados en férreas clasificaciones, en la que cada cual va por libre autogestionando su propio itinerario, aprendiendo de aquí y de allá, saltando de un tema a otro, enfrascado en sus procesos personales, acoplado en una red de relaciones que se activan o desactivan según convenga. ¡Ojo!

La inteligencia artificial es un camino de ida y vuelta. Por una parte aspira a emular la inteligencia humana y replicarla en sistemas artificiales. Ese camino va de la inteligencia humana a la artificial. Pero, por otra, aspira a construir máquinas cuyos funcionamientos

contribuyan a explicar qué es la inteligencia humana. Ese camino va de la artificial a la humana. Hacer la artificial para comprender la humana. Algoritmos y computación al servicio de la neurociencia para responder las grandes preguntas: ¿cómo funciona la inteligencia?, ¿qué es el aprendizaje?, ¿cómo emergen las capacidades cognitivas del cerebro humano?

Para responder estas preguntas, para ir de la artificial a la humana, no da igual el tipo de inteligencias artificiales que se construyan. No da igual el tipo de experimentos que se hagan. Más allá de la utilidad práctica que alcancen a tener, más allá de los problemas concretos que puedan resolver, hay que construirlas eligiendo un nivel de abstracción adecuado.

Las inteligencias artificiales conexionistas están consiguiendo resultados útiles y la industria se ha lanzado a desarrollar aplicaciones (más o menos) prácticas, usables, vendibles. La ingeniería ha tomado la delantera a la ciencia. Es el triunfo de la ingeniería. Máquinas eficaces, sí. Pero ¿responden a las preguntas fundamentales?

Una parte de la comunidad científica alerta del cambio que supone pasar de una ciencia basada en teoría a una ciencia basada en datos. ¿Qué es lo que va a guiar a la ciencia? ¿La búsqueda de conocimiento o la disponibilidad de datos? ¿Van a ser los datos los que generen las hipótesis? El pensamiento teórico no desaparece, claro está, pero se ve desafiado por las nuevas herramientas.

Para crear una inteligencia artificial que prediga el tiempo se precisa de las meteorólogas y para crear una que haga reconocimiento de voz se precisa de las lingüistas. Cierto. Pero ¿cómo van a seguir aprendiendo las meteorólogas y las lingüistas? Las herramientas de inteligencia artificial pueden producir mejores resultados que los humanos. Eso se dice, por ejemplo, de AlphaFold, entrenado para predecir la estructura de las proteínas. Pero, si no se pueden reproducir, verificar y explicar sus resultados de forma fiable y plena, ¿qué tipo de conocimiento se genera? ¿Cómo sigue aprendiendo el humano?

En el extremo de estas alertas está la suposición de que el aprendizaje automático en realidad no puede producir nada nuevo y, por tanto, es incapaz de, en sentido estricto, descubrir nada. E incluso se aventura la hipótesis de que a futuro solo las inteligencias artificiales podrán entenderse y aprender entre sí, en procesos opacos a la comprensión humana.

Estas hipótesis tan drásticas ¿son lamentos de los sistemas de autoridad tradicionales establecidos, que gimen como un rey destronado? ¿O abren un debate necesario y pertinente? En temas de salud pública, por ejemplo, el medio camino entre aceptar el buenismo y la objetividad de los datos masivos o verlos como una fuente de información llena de ruido y de sesgos señala las dificultades con las que se enfrenta la investigación: la objetividad de los datos, la transparencia de los métodos y la interpretación y replicabilidad de los resultados.

Funcionar manejando datos que no se comprenden puede ser pan para hoy y hambre para mañana. No hay nada en contra de manejar datos que no se comprenden, siempre y cuando sea para aprender a extraer información significativa de ese aparente caos. Los datos por sí solos no ayudan a descubrir qué está pasando si no hay unas hipótesis sobre lo que se está buscando. Hasta ahora, la ciencia ha necesitado teoría, leyes. Así, al menos, es como ha funcionado desde Newton. Una aproximación de tipo conductista, sin poder explicativo, no es válida.

Pero otra parte de la ciencia critica esta crítica: tu crítica está anclada en el paradigma de la ciencia teórica, en la vieja ciencia. Ahora, la avalancha de datos disponibles ha modificado el paradigma. Estamos en una ciencia de datos. Quizás haya que sacrificar un poco de interpretabilidad, pero, sin miedo, abracemos el cambio de paradigma y ¡hagamos ciencia de datos!

Desde la propia inteligencia artificial la cuestión de la inexplicabilidad es tan crítica que más allá de los enfoques simbólico o conexionista se están estudiando enfoques híbridos que mezclen lo mejor de ambos, dando la explicabilidad simbólica sin perder las ventajas del conexionismo.

Uno de los enfoques para construir redes neuronales explicables es el de la inteligencia artificial neurosimbólica, que combina el razonamiento simbólico con el aprendizaje automático, defendiendo que la única manera de manipular de forma confiable el conocimiento abstracto es la simbólica, pero reconociendo las ventajas conexionistas, especialmente el reconocimiento de patrones. Estos sistemas podrían reunir lo mejor de los dos mundos, requerir menos datos de entrenamiento y ser menos sensibles al ruido en ellos. Fusionar lógica y aprendizaje. El objetivo: convertir los modelos de redes neuronales artificiales de caja negra en modelos que se puedan entender

desde la lógica, pero no necesariamente desde la lógica del razonamiento humano. Los humanos no son confiables como para describir el conocimiento sobre el mundo. Las máquinas pueden hacerlo mejor si consiguen comprender, verdaderamente comprender, el lenguaje natural.

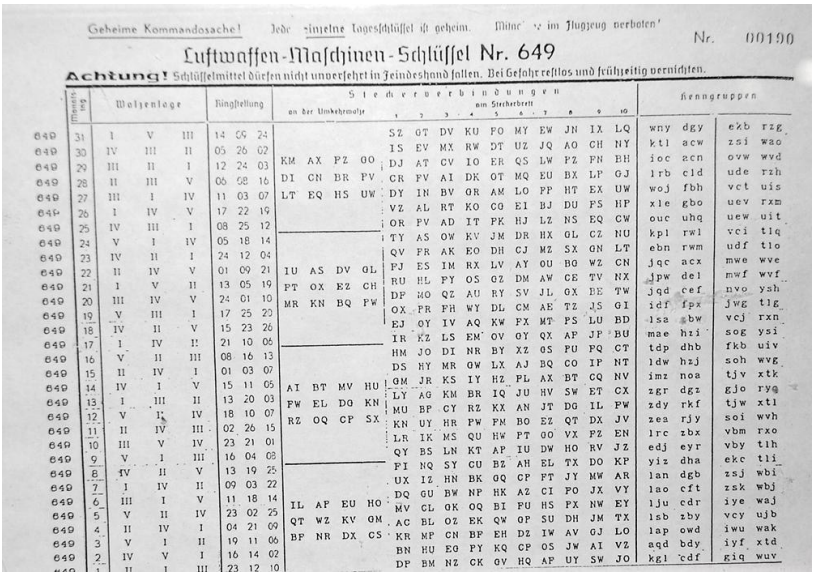
Tradúceme

Una de las aplicaciones del procesamiento del lenguaje natural es la traducción automática de texto.

Después de la Segunda Guerra Mundial, en la época de la Guerra Fría, la traducción automática de texto se convirtió en un área de investigación estratégica tanto para Estados Unidos como para la Unión de Repúblicas Socialistas Soviéticas (URSS), las dos superpotencias enfrentadas a nivel político, económico, social, ideológico, militar y propagandístico. Ambas ansiaban poder traducir automáticamente documentos publicados en el otro bloque o interceptados con mayor o menor grado de espionaje. Obviamente, el objetivo no era traducir poesía, sino todo lo que tuviera que ver con política, matemáticas, ciencia, tecnología e industria.

En Estados Unidos escaseaban los traductores de ruso y había mucho texto ruso por traducir. La posibilidad de la traducción automática había sido alentada por los éxitos cosechados durante la Segunda Guerra Mundial en materia de criptografía. Durante la guerra, el bloque de los aliados había conseguido romper el código de la máquina Enigma. Esta era una máquina de cifrado electromecánica (no informática) que se había inventado en Alemania y usaban las empresas para sus comunicaciones en clave con el fin de zafarse del espionaje industrial.

Durante la Segunda Guerra Mundial el ejército alemán la mejoró y la utilizaba para las comunicaciones militares. Se consideraba indescifrable. Los mensajes, escritos originariamente en alemán, se encriptaban con una máquina Enigma y con la ayuda de unas listas



Geheime Kommandosache! Jede einzelne Tagesschlüssel ist geheim. Mitnehmen im Flugzeug verboten!

Nr. 00190

Luftwaffen-Maschinen-Schlüssel Nr. 649

Achtung! Schlüsselmittel dürfen nicht unversehrt in Feindeshand fallen. Bei Gefahr restlos und frühzeitig vernichten.

Monats tag	Walzenlage			Ringstellung			an der Umkehrwalze	Steckerverbindungen am Steckerbrett										Kenngruppen	
								1	2	3	4	5	6	7	8	9	10		
649	31	I	V	III	14	09	24		SZ	GT	DV	KU	FO	MY	EW	JN	IX	LQ	wny dgy	ezb rzg
649	30	IV	III	II	05	26	02		IS	EV	MX	RW	DT	UZ	JQ	AO	CH	NY	ktl acw	zsi wao
649	29	III	II	I	12	24	03	KM AX PZ OO	DJ	AT	CV	IO	ER	QS	LW	PZ	FN	BH	joc zcn	ovw wvd
649	28	II	III	V	06	08	16	DI GN BR FV	CR	FV	AI	DK	OT	MQ	EU	BX	LP	GJ	lrb cld	ude rzh
649	27	III	I	IV	11	03	07	LT EQ HS UW	DY	IN	BV	OR	AM	LO	FP	HT	EX	UW	woj fbh	vct uis
649	26	I	IV	V	17	22	19		VZ	AL	RT	KO	CG	EI	BJ	DU	PS	HP	xle gbo	uev rxm
649	25	IV	III	I	08	25	12		OR	FV	AD	IT	PK	HJ	LZ	NS	EQ	CW	ouc uhq	uew uit
649	24	V	I	IV	05	18	14		TY	AS	OW	KV	JM	DR	HX	GL	CZ	NU	kpl rwl	vci tlq
649	23	IV	II	I	24	12	04		QV	FR	AK	EO	DH	CJ	MZ	SX	GN	LT	ebn rwm	udf tlo
649	22	II	IV	V	01	09	21	IU AS DV GL	FJ	ES	IM	RX	LV	AY	OU	BQ	WZ	CN	jqc acx	mwe wve
649	21	I	V	II	13	05	19	PT OX EZ CH	RU	HL	FY	OS	QZ	DM	AW	CE	TV	NX	jpw del	mwf wvf
649	20	III	IV	V	24	01	10	MR KN BQ FW	DF	MO	QZ	AU	BY	SV	JL	QX	BE	TW	jqd cef	nvo ysh
649	19	V	III	I	17	25	23		OX	PR	FH	WY	DL	CM	AE	TZ	JS	GI	idf fpx	jwg tlg
649	18	IV	II	V	15	23	26		EJ	OY	IV	AQ	KW	FX	MT	PS	LU	BD	lsa sbw	vcj rxn
649	17	I	IV	II	21	10	06		IR	KZ	LS	EM	OV	OY	QX	AP	JP	BU	mae hzi	sog ysi
649	16	V	II	III	08	16	13		HM	JO	DI	NR	BY	XZ	GS	PU	PQ	CT	tdp dhb	fkb uiv
649	15	II	IV	I	01	03	07		DS	HY	MR	GW	LX	AJ	BQ	CO	IP	NT	ldw hzj	soh wvg
649	14	IV	I	V	15	11	05	AI BT MV HU	QM	JR	KS	IY	HZ	PL	AX	BT	GQ	NV	imz noa	tjv xtk
649	13	I	III	II	13	20	03	FW EL DG KN	LY	AG	KM	BR	IQ	JU	HV	SW	ET	CX	zgr dgz	gjo ryq
649	12	V	II	IV	18	10	07	RZ OQ CP SX	MU	BP	CY	RZ	KX	AN	JT	DG	IL	PW	zdy rkf	tjw xtl
649	11	II	IV	III	02	26	15		KN	UY	HR	PW	FM	BO	EZ	QT	DX	JV	zea rjy	soi wvh
649	10	III	V	IV	23	21	01		LR	IK	MS	QU	HW	PT	GO	VX	PZ	EN	lrc zbx	vbm rxo
649	9	V	I	III	16	04	08		QY	BS	LN	KT	AP	IU	DW	HO	RV	JZ	edj eyr	vby tlh
649	8	IV	II	V	13	19	25		FI	NQ	SY	GU	BZ	AH	EL	TX	DO	KP	yiz dha	ekc tli
649	7	I	IV	II	09	03	22		UX	IZ	HN	BK	OQ	CP	FT	JY	MW	AR	lan dgb	zsj wbi
649	6	III	I	V	11	18	14		DQ	GU	BW	NF	HK	AZ	CI	PO	JX	VY	lao cft	zsk wbj
649	5	V	II	IV	23	02	25	IL AF EU HO	MV	CL	GK	OQ	BI	FU	HS	PX	NW	EY	ljy cdr	iye waj
649	4	II	IV	I	04	21	09	QT WZ KV GM	AC	BL	OZ	EK	QW	GP	SU	DH	JM	TX	lsb zby	vcy ujb
649	3	V	I	II	19	11	06	BF NR DX CS	KR	MP	CN	BF	EH	DZ	IW	AV	GJ	LO	lap owd	iwu wak
649	2	IV	V	I	16	14	02		BN	HU	EQ	FY	KQ	CP	OS	JW	AI	VZ	aqd bdy	iyf xtd
649	1	II	I	III	23	12	10		DP	BM	NZ	CK	OV	HQ	AP	UY	SW	JO	kgl cdf	giq wuv

Imagen 14. Lista mensual de claves número 649 para la Enigma de las Fuerzas Aéreas alemanas, incluidos los ajustes del reflector reconfigurable, que solo cambian una vez cada ocho días. (Fuente: https://en.wikipedia.org/wiki/Enigma_machine).

de claves. El mensaje se transmitía encriptado y en el destino se desencriptaba con otra Enigma y las mismas listas de claves.

Al principio de la guerra, en Inglaterra, en la mansión victoriana Betchley Park, los aliados camuflaron una instalación secreta para que matemáticos y mayoritariamente matemáticas, pues el 75 % de las criptógrafas eran mujeres, junto con lingüistas, jugadores de ajedrez y *frikies* de los crucigramas trabajaran para descifrar los imposibles códigos de las Enigma alemanas. Diez mil personas trabajando durante dieciocho meses hasta que, en colaboración con colegas de Polonia, lo consiguieron.

Si se había conseguido desencriptar ese código endiablado, ¿no se iba a poder traducir del ruso al inglés? El razonamiento era: en lugar de pensar que el texto está escrito en ruso, pensemos que está escrito en inglés y que luego se ha encriptado con símbolos extraños (el ruso). Se trata «simplemente» de romper el código. De repente, en las potencias occidentales y, por los mismos motivos, en la Unión

Soviética empezó a haber mucho dinero para investigar en traducción automática, siguiendo un halo de inspiración cibernética.

En 1949 se publicó el memorándum Weaver, un documento que se tomaba en serio la posibilidad de utilizar el reciente invento de las computadoras para realizar traducción automática y establecía cuatro hipótesis de trabajo para superar el enfoque simplista de la traducción palabra por palabra.

La primera hipótesis era que el problema de la polisemia (pluralidad de significados distintos para una única palabra) se podría resolver si se tenía en cuenta el contexto, entendiendo por contexto unas cuantas palabras de las que iban antes y otras cuantas de las que iban después de la palabra polisémica.

La segunda era que el lenguaje es lógico, contiene lógica, y por tanto el problema de la traducción tiene una solución formal (lógica). Traducir es un problema de lógica, de modo que se puede resolver con algoritmos lógicos ejecutados en computadoras.

La tercera era que todo lenguaje tiene propiedades estadísticas: hay frecuencias de letras, intervalos entre letras, combinaciones de letras, patrones de letras… Esta hipótesis planteaba aplicar a la traducción un enfoque criptográfico. Venía a decir que traducir es desencriptar.

Y la cuarta era que hay invariantes lingüísticos que son comunes a todas las lenguas. En la medida en que todas las lenguas han sido inventadas y desarrolladas por humanos, y todos los humanos tienen cerebros de complejidad similar, cabe esperar que el lenguaje haya surgido de manera similar en diferentes lugares y épocas, por lo que se puede suponer que, aunque las lenguas presenten muchas diferencias superficiales, también tendrán aspectos básicos comunes, si bien pueden estar escondidos y no ser evidentes a primera vista.

Aunque cada una de estas hipótesis abría sus propios recorridos lingüísticos, computacionales y matemáticos, quedaban trazadas las dos líneas de trabajo principales: los métodos lingüísticos, basados en reglas, y los métodos estadísticos, basados en corpus.

En la traducción automática basada en reglas, personas que conocen la lengua de origen y la lengua de destino escriben diccionarios y reglas de transformación estructurales. Las informáticas programan los motores que usarán esos diccionarios y aplicarán esas reglas de transformación. El motor analiza lingüísticamente el texto original a distintos niveles más o menos profundos (morfológico, sintáctico,

semántico…), establece los equivalentes en la lengua de destino y produce la traducción siguiendo unas reglas de generación.

En la traducción automática basada en corpus no hay análisis de la lengua. Se usan amplias recopilaciones de textos escritos en la lengua de origen con centenares de miles o incluso millones de frases ya traducidas o que traducen expresamente las traductoras. Las informáticas crean programas de ordenador que aprenden cuál es, estadísticamente, la mejor traducción para cada frase. Son sistemas probabilísticos.

Tras la publicación del memorándum Weaver, se pusieron manos a la obra y en el Massachusetts Institute of Technology (MIT), bajo la hipótesis de que las lenguas son lógicas, comenzaron a buscar leyes que sentaran las bases de una ciencia explicativa del lenguaje natural. Hacían lingüística teórica. Buscaban una gramática asentada en formulaciones lógico-matemáticas. Y así se puso en marcha el ambicioso proyecto Fully Automatic High Quality Traslation (FAHQT).

Por su parte, la RAND Corporation, desde su laboratorio de ideas, abordó un enfoque empírico basado en reglas codificadas a mano, por lingüistas, para traducir palabras y frases apoyándose en el análisis de la estructura gramatical de la oración y en el uso de diccionarios bilingües. Se programaron motores que funcionaban medio bien para lenguas con estructuras gramaticales muy definidas, pero no manejaban bien las indefiniciones ni las excepciones.

A la vuelta de veinte años de desarrollos desde aproximaciones diversas, en 1964 las principales fuentes de financiación estadounidenses, es decir, las agencias militares y de inteligencia, instaron a la National Science Foundation a que constituyera un comité asesor que estudiase el estado de la cuestión, y así se formó el Automatic Language Processing Advisory Committee (ALPAC). El comité recibió el encargo de evaluar el progreso de la lingüística computacional en general y de la traducción automática en particular.

Después de dos años de estudio, publicó un informe demoledor: no se había conseguido ninguna traducción automática de calidad y no había ninguna perspectiva inmediata de conseguirla. Continuar investigando en traducción automática era perder el tiempo.

El informe ALPAC detectaba tres problemas insalvables: la calidad, la velocidad y el coste. La calidad no alcanzaba unos mínimos. Debido a esa carencia, el lector tardaba el doble de tiempo en leer y comprender un texto traducido automáticamente que uno traducido

por humanos. En cuanto a la velocidad, con los ordenadores y los métodos de la época, una traducción de 50 páginas requería como mínimo 15 días para completarse. A todas luces era mucho más barato pagar a traductores humanos o enseñar ruso a los científicos. Era un informe práctico, centrado en las necesidades gubernamentales y militares de Estados Unidos de analizar y digitalizar documentos en ruso y, aunque reconocía que se habían realizado avances teóricos en cuanto al procesamiento del lenguaje natural, recomendaba retirar la financiación para la traducción automática y reconducirla a la investigación en asistentes a la traducción que dotaran a la traducción humana de mayor velocidad. El informe proponía reorientar la investigación hacia la traducción asistida: diccionarios *online*, textos predictivos, bases de datos lingüísticas… La computadora podía ser una muy buena herramienta como asistente para el humano en la traducción, pero siempre, siempre, siempre sería necesaria la posedición (humana).

A pesar de este viraje, Estados Unidos no era el único implicado en el asunto. Estaba la Unión Soviética. Estaba Japón. Estaban los Estados bilingües, como Canadá, que prototipó un sistema basado en reglas, el Météo, para traducir los partes meteorológicos del francés al inglés y viceversa. Con un vocabulario restringido y sintaxis limitada, funcionaba muy bien. Se mantuvo operativo hasta 2001.

También estaba la recién constituida Comunidad Económica Europea, la Europa de los seis, que desde su origen, cuando el Tratado de Roma entró en vigor en 1958, asumió el multilingüismo y estableció cuatro idiomas oficiales: alemán, francés, italiano y neerlandés. Aunque al principio se utilizaba más la interpretación, imprescindible en las primeras negociaciones, la traducción también era necesaria, especialmente para textos científicos, técnicos, legales y administrativos. Cada institución dedicaba entre el 35 % y el 65 % de su presupuesto a las traducciones.

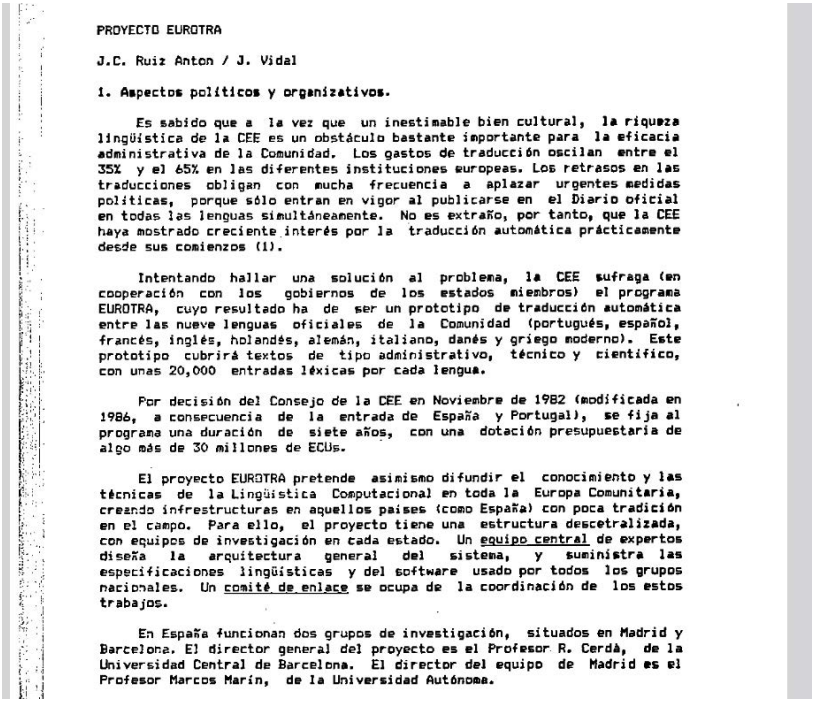
En el año 1978 se puso en marcha el proyecto Eurotra para los idiomas oficiales, que en ese momento eran alemán, danés, francés, neerlandés, inglés, italiano y poco más tarde español, griego y portugués.

Eurotra no llegó a funcionar. Pronto se asumió que se trataba más de un proyecto de investigación que de desarrollo de traductores automáticos operativos. Sin embargo, es un buen ejemplo del peso que tienen las decisiones políticas: la ciudadanía tiene derecho a leer los documentos de la Comunidad en su propio idioma cueste lo que cueste.

Además, pretendía difundir el conocimiento y las técnicas de la lingüística computacional por toda la Europa comunitaria. Había que diseminar saberes e infraestructuras, así que se concibió como un proyecto descentralizado.

Nunca funcionó, porque se tuvo que enfrentar a la complejidad: para traducir cualquier frase se podía aplicar un gran número de reglas gramaticales lo que producía miles de derivaciones que ralentizaban el proceso de traducción. Sin embargo, fue un fracaso muy exitoso. Generó y diseminó mucho conocimiento, especialmente en los países del sur: España, Grecia, Italia y Portugal.

No solo estaban Estados Unidos, la Unión Soviética y Japón. No solo estaban la geopolítica y los mercados. También estaba el amor a la lengua propia. Estaban las comunidades. Y estaba el software libre.



PROYECTO EUROTRA

J.C. Ruiz Anton / J. Vidal

1. Aspectos politicos y organizativos.

Es sabido que a la vez que un inestimable bien cultural, la riqueza lingüística de la CEE es un obstáculo bastante importante para la eficacia administrativa de la Comunidad. Los gastos de traducción oscilan entre el 35% y el 65% en las diferentes instituciones europeas. Los retrasos en las traducciones obligan con mucha frecuencia a aplazar urgentes medidas politicas, porque sólo entran en vigor al publicarse en el Diario oficial en todas las lenguas simultáneamente. No es extraño, por tanto, que la CEE haya mostrado creciente interés por la traducción automática prácticamente desde sus comienzos (1).

Intentando hallar una solución al problema, la CEE sufraga (en cooperación con los gobiernos de los estados miembros) el programa EUROTRA, cuyo resultado ha de ser un prototipo de traducción automática entre las nueve lenguas oficiales de la Comunidad (portugués, español, francés, inglés, holandés, alemán, italiano, danés y griego moderno). Este prototipo cubrirá textos de tipo administrativo, técnico y científico, con unas 20,000 entradas léxicas por cada lengua.

Por decisión del Consejo de la CEE en Noviembre de 1982 (modificada en 1986, a consecuencia de la entrada de España y Portugal), se fija al programa una duración de siete años, con una dotación presupuestaria de algo más de 30 millones de ECUs.

El proyecto EUROTRA pretende asimismo difundir el conocimiento y las técnicas de la Lingüistica Computacional en toda la Europa Comunitaria, creando infrestructuras en aquellos paises (como España) con poca tradición en el campo. Para ello, el proyecto tiene una estructura descetralizada, con equipos de investigación en cada estado. Un equipo central de expertos diseña la arquitectura general del sistema, y suministra las especificaciones lingüísticas y del software usado por todos los grupos nacionales. Un comité de enlace se ocupa de la coordinación de los estos trabajos.

En España funcionan dos grupos de investigación, situados en Madrid y Barcelona. El director general del proyecto es el Profesor R. Cerdà, de la Universidad Central de Barcelona. El director del equipo de Madrid es el Profesor Marcos Marín, de la Universidad Autónoma.

Imagen 15. Documento del proyecto Eurotra.

Consérvame

En el año 2004, en el Departamento de Lenguajes y Sistemas Informáticos de la Universitat d'Alacant (Alicante), con dinero público del gobierno de España y de la Generalitat de Catalunya, el grupo de investigación Transducens empezó a trabajar en el proyecto Apertium. Se quería hacer una plataforma que ofreciera traductores automáticos entre distintas lenguas, empezando por el par castellano-catalán.

En la página oficial de Apertium podemos leer:

> Triste pero cierto: en nuestro mundo, las lenguas, como el entorno natural, están desapareciendo. En los últimos doscientos años, más o menos, se han extinguido lenguas, y muchas están amenazadas.

Se necesitaba un traductor para lenguas amenazadas, minoritarias o sin Estado. Había que proteger las lenguas en peligro de extinción y evitar que otras llegasen a esa situación. El grupo de promotores del proyecto tenía un objetivo muy claro: desarrollar de forma colaborativa un traductor automático bajo los siguientes principios:

1. Ofrecer a todo el mundo acceso gratuito e ilimitado a las mejores tecnologías de traducción automática posibles.

2. Mantener una plataforma modular, documentada y abierta para la traducción automática y otras tareas de procesamiento humano del lenguaje.

3. Favorecer el intercambio y la reutilización de los datos lingüísticos existentes.

4. Facilitar la integración con otras tecnologías libres (software libre).

5. Garantizar absolutamente la reproducibilidad de la investigación en la traducción automática y en el procesamiento del lenguaje natural. Consecuentemente, su código se distribuiría con una licencia de software libre, en concreto con la licencia GPL.

Por esas fechas, debido a unas computadoras más potentes y a la disponibilidad de conjuntos de datos abundantes —en parte gracias a Internet—, había cobrado auge la traducción automática basada en corpus, es decir, estadística. Con ello se daba un giro de ciento ochenta grados a las estrategias de traducción basadas en reglas. Adiós a los diccionarios y a las reglas. La idea era que, en lugar de que los lingüistas definieran las reglas, el ordenador podía aprender por sí mismo las relaciones existentes entre dos lenguas si le dábamos la oportunidad de aprender con datos suficientes. Este aprendizaje de enfoque conexionista, que todavía no usaba redes neuronales, utilizaba modelos probabilísticos que seleccionaban la salida (la traducción) más probable para la entrada (el texto a traducir), calculando las probabilidades mediante el procesamiento de grandes corpus de entrenamiento paralelos en ambos idiomas.

Aunque desde algunos sectores de la lingüística teórica se señalaba que arrancar traducciones a fuerza de exprimir la potencia de cálculo estadístico de las máquinas no es ciencia, el caso es que eso funcionaba, aunque no sin problemas. Pero, claro, la traducción basada en reglas también tenía problemas.

Los sistemas estadísticos, entrenados con corpus, daban traducciones más fluidas, más naturales, mientras que los basados en reglas entregaban salidas más mecánicas, menos naturales, aunque muchas veces más exactas. Esto es así porque, en cierto modo, los estadísticos ofrecían como salida frases ya hechas, entresacadas de los corpus ya traducidos por humanos.

Pero presentaban el problema de que, para entrenarlos, se necesitaban millones de pares de oraciones representativas de todos los dominios de la lengua, de todas las temáticas, registros, etcétera, en los que se quisiera aplicar la traducción. Tal cantidad de datos no iba a estar disponible para las lenguas amenazadas, minoritarias o sin Estado.

Translators

The following 49 pairs have released versions and are considered to be stable:

- Spanish ⇆ Catalan (es-ca)
- Spanish ← Romanian (es-ro)
- French ⇆ Catalan (fra-cat)
- Occitan ⇆ Catalan (oc-ca)
- English ⇆ Galician (en-gl)
- Swedish ⇆ Danish (swe-dan)
- Macedonian → English (mk-en)
- Afrikaans ⇆ Dutch (af-nl)
- Indonesian ⇆ Malaysian (id-ms)
- Icelandic ⇆ Swedish (is-sv)
- Occitan ⇆ Spanish (oc-es)
- Spanish ⇆ Portuguese (es-pt)
- English ⇆ Catalan (en-ca)
- English ⇆ Spanish (en-es)
- Sardinian ← Italian (ita-srd)
- Sardinian ← Catalan (cat-srd)
- English ⇆ Esperanto (en-eo)
- Spanish → Asturian (spa-ast)
- Catalan ← Italian (ca-it)
- Maltese → Arabic (mlt-ara)
- Serbo-Croatian ⇆ Slovenian (hbs-slv)
- Danish ⇆ Norwegian (dan-nor)
- Spanish ⇆ Galician (es-gl)
- French ⇆ Spanish (fr-es)
- French → Occitan (fra-oci)
- Esperanto ← Spanish (eo-es)
- Welsh → English (cy-en)
- Breton → French (br-fr)
- Icelandic → English (isl-eng)
- Esperanto ← Catalan (eo-ca)
- North Sámi → Norwegian (sme-nob)
- Crimean Tatar → Turkish (crh-tur)
- Serbo-Croatian → Macedonian (hbs-mkd)
- Serbo-Croatian → English (hbs-eng)
- Portuguese ⇆ Catalan (pt-ca)
- Portuguese ⇆ Galician (pt-gl)
- Basque → Spanish (eu-es)
- Norwegian Nynorsk ⇆ Bokmål (nno-nob)
- Macedonian ⇆ Bulgarian (mk-bg)
- Esperanto ← French (eo-fr)
- Basque → English (eu-en)
- Spanish ⇆ Aragonese (spa-arg)
- Kazakh ⇆ Tatar (kaz-tat)
- Urdu ⇆ Hindi (urd-hin)
- Aragonese ⇆ Catalan (arg-cat)
- Swedish ⇆ Norwegian (swe-nor)
- Belarusian ⇆ Russian (bel-rus)
- Russian ⇆ Ukrainian (rus-ukr)
- Polish → Silesian (pol-szl)

Imagen 16. Pares de lenguas traducibles con Apertium. (Fuente: https://web.archive.org/web/20190924155344/http://wiki.apertium.org/wiki/Main_Page).

Había todavía otro gran problema: el modelo estadístico es difícil de mantener. Con muchos datos da buenos resultados rápidamente, pero actualizar su aprendizaje es muy costoso.

Y otro más: su estilo de programación informática no se presta bien al trabajo colaborativo.

Es por todo ello que Apertium se decantó por construir un sistema basado en reglas. Un sistema basado en reglas permitiría organizar mejor la cooperación de lingüistas, traductoras e informáticas en pro de la generación y el mantenimiento de la plataforma, pues las colaboradoras podrían ir realizando el trabajo incremental de mantener las reglas y agregar nuevas. Y el resultado sería suficientemente bueno para el objetivo tecnosocial del proyecto, ya que la mayoría de pares de lenguas iban a ser similares entre sí.

Apertium es un ejemplo muy claro de que detrás de todo proyecto tecnológico, comercial o no, hay una visión sobre cómo se insertará, cómo hará simbiosis con los entornos humanos concretos. Muestra cómo en la mayoría de los casos hay un amplio margen de decisiones tecnológicas posibles. Y cómo cada una de las decisiones decanta el proyecto hacia donde se quiera deslizar. Las tecnologías no son lentejas, que si las quieres las tomas y si no las dejas. Hay margen. Todo se puede hacer de maneras muy diferentes.

Con el tiempo se han ido desarrollando muchos traductores para distintos pares de lenguas, pero no es solo eso. Con su código libre disponible y descargable en GitHub, donde se reportan treinta y dos personas participantes en el desarrollo y el mantenimiento, es un proyecto vivo que, fiel al principio de la reproducibilidad de la investigación, ofrece un andamiaje para que cualquiera con los conocimientos suficientes tenga la posibilidad de desarrollar más traducciones, para que más lenguas, amadas y amenazadas, sean traducibles. Para que el número 2.450, que son las lenguas de todo el mundo que se calcula que están en peligro de extinción, no se incremente.

Mucha gente usa Apertium sin saberlo. Por ejemplo, Softcatalà —un grupo de activistas de la lengua catalana en la computación y en Internet que se dedica, entre otras cosas, a traducir al catalán las interfaces de muchos programas de ordenador, como LibreOffice o Mozilla— ofrece en su web, junto con otras herramientas, como un corrector, diccionarios o un transcriptor de audio o vídeo a texto

(todo ello usando únicamente software libre), una instancia de este traductor con el añadido de una privacidad de datos total.

En la actualidad, Softcatalà ofrece un traductor neuronal, ya operativo, para el par de lenguas catalán-inglés. Cuando, en su web, se pide una traducción para ese par, la usuaria decide si quiere usar su traductor neuronal o Apertium.

Desde hace unos diez o quince años, la traducción automática se ha desarrollado utilizando redes neuronales artificiales. Al igual que los estadísticos, los traductores neuronales se entrenan utilizando grandes corpus paralelos. La especificidad es que lo hacen usando redes neuronales artificiales. Más que generar traducciones a partir del análisis del texto, lo que hacen es buscar traducciones previas hechas por humanos y seleccionar y reutilizar las más probables. En 2016 el traductor de Google dejó de ser estadístico y pasó a ser neuronal.

El enfoque neuronal da buenos resultados, gestiona mejor el contexto de las palabras, pero también tiene problemas. Para el entrenamiento requiere hardware especializado y potente que no está al alcance de pequeñas empresas o traductoras autónomas. Tiene un punto de no explicabilidad y en las frases largas cuesta hacer el *match* entre las palabras del texto original y las de la traducción. Si se encuentra con una palabra nueva, que no aprendió durante su aprendizaje, la sustituye por otra de manera muchas veces incoherente. Como trabaja a nivel de subpalabra, puede inventarse palabras, como por ejemplo «mecanaje». También puede inventarse texto que no estaba en el original e insertarlo con una corrección gramatical absoluta, lo que se conoce como alucinación. O puede eliminarlo si lo considera poco probable. Y da como salida textos muy fluidos, lo cual en principio supone una ventaja, pero, como veremos más adelante, también puede ser un problema.

Además, si traduce mal una palabra no se puede reparar directamente. No se puede modificar una red neuronal artificial ya creada, que tiene millones de conexiones, cada una con su número w. Lo que se puede hacer es corregir los textos con los que se ha entrenado a la red y volverla a entrenar. Ese proceso es costoso.

En agosto del 2020, año de la pandemia, Softcatalà presentó un traductor automático inglés-catalán y viceversa desarrollado por el propio grupo basándose en tecnología de red neuronal artificial. Entrenaron al traductor con 4,5 millones de frases traducidas por humanos tomadas de distintas fuentes, como por ejemplo la Viquipèdia

(Wikipedia en lengua catalana), y se esforzaron para que esas traducciones fueran de calidad. Su hipótesis de trabajo era que para la traducción automática entre lenguas lejanas la mejor tecnología es la neuronal. En sus propias palabras: «Pensamos que para lenguas próximas, como castellano y catalán, los sistemas de reglas (como por ejemplo la Apertium, que ya usamos) funcionan mejor que los neuronales. No pensamos que se pueda hacer algo significativamente mejor de lo que ya tenemos. Por otro lado, ahora mismo no tenemos el hardware necesario para asumir la carga que representaría».

Softcatalà es una comunidad que autodetermina las tecnologías convenientes para cada problema tecnosocial que se trae entre manos. Tiene conocimientos y tiene posicionamiento. Qué duda cabe de que un traductor automático, utilice la tecnología que utilice, es una inteligencia artificial. Sin embargo, en su web la narrativa sobre las promesas de la inteligencia artificial no tiene ninguna relevancia. No necesitan celofán. No necesitan *marketing*. En cambio, sí la tiene la defensa de la lengua propia, la lucha contra la privatización del conocimiento y la defensa de los recursos comunes. Por eso publican los modelos neuronales y los corpus que utilizan. Hay inteligencias artificiales que no necesitan exhibirse.

El caso de los traductores automáticos es bueno para ilustrar cómo una transformación tecnológica salpica (y conecta) ámbitos muy distintos: desde lo laboral a lo filosófico, desde lo político a lo científico, desde lo matemático a lo lingüístico.

Para las personas que trabajan en la traducción, la mejora de la calidad en los traductores automáticos cambia las reglas del juego. Cuando entra la producción automatizada, en serie, ya no se puede seguir practicando un oficio con métodos artesanales. La transformación profesional se presenta, en principio, amenazante: pérdida de trabajos y desvalorización de la actividad, es decir, precarización. Y más precarización cuando el sector no está organizado, carece de alianzas transversales, no puede crear conflicto, no tiene voz. No tiene, pero ¿podría tenerla?

Por ejemplo, respecto a la calidad de un traductor automático. ¿Quién la evalúa y cómo se evalúa?

El concepto de la calidad de una traducción es complejo y está sujeto a debates. La calidad depende del contexto y el contexto cambia, por lo que es difícil establecer un modelo universal, un estándar. Cierto que hay errores, pero para identificar un error antes hay que saber

cómo se debe o se quiere traducir. Hay que establecer una norma, que puede depender del contexto. Hay muchas normas. ¿Se busca una equivalencia formal? ¿O se busca una equivalencia semántica? ¿El lenguaje al que se traduce requiere cambios de significado? ¿Qué nivel de calidad se espera? ¿Suficiente? ¿Alta? ¿Es aceptable que el traductor cometa errores que no afectan a la comprensión del texto (por ejemplo, la falta de un artículo)? ¿Cuánto presupuesto económico hay disponible para esa traducción?

Imaginemos, salvando todas las dificultades y haciendo un esfuerzo de simplificación, que un traductor tiene un 98 % de calidad. De cada cien palabras, solo hay dos que son erróneas. Eso está muy bien. Parece que la posedición va a suponer poco trabajo, solo cambiar dos palabras de cada cien. Parece fácil, pero ¿cuánto esfuerzo cognitivo, conocimiento, atención, tiempo y cansancio conlleva detectar cuáles son esas dos palabras?

Esa posedición ¿se hace mejor —con menos tiempo o con menos cansancio— con un traductor basado en reglas? ¿O con uno neuronal? ¿O depende de los idiomas? ¿O depende de la norma elegida? ¿O de la temática? ¿Se posedita mejor un texto muy fluido con pocos errores agazapados que otro menos fluido con más errores evidentes? Todo eso es un conocimiento crítico y valioso que solo puede surgir de las traductoras.

Para la traducción profesional, ¿es útil la traducción automática? ¿Es mejor la asistida? ¿O hay que olvidarse de los traductores genéricos, tipo DeepL, y empezar a pensar en que cada traductora tenga su propio programa traductor, instalado en su ordenador local y entrenado con el corpus de sus propias traducciones? Esto no es sencillo, pero empieza a ser viable. Todo el software necesario para entrenar y utilizar sistemas neuronales se distribuye gratuitamente bajo licencias libres. Existen andamiajes, como Marian u OpenNMT, con todo lo necesario para tener un traductor automático personalizado en el ordenador personal, con el beneficio de la privacidad total. Es cierto que para el entrenamiento (que no para el uso) se requieren ordenadores dotados con procesadores GPU potentes, pero estos procesadores están bajando de precio y los ordenadores orientados al videojuego los tienen incorporados. También es cierto que se necesitan conocimientos informáticos. El acceso a todo esto es más fácil para empresas de servicios lingüísticos grandes que para una traductora autónoma. Pero difícil no significa imposible. Especialmente, si se construye el propósito tecnosocial de la acción.

Por otra parte, si la traducción es automática, ¿qué se paga cuando se paga por una traducción? ¿La traductora debe empezar a considerarse más bien una poseditora, un evaluador humano que valida o certifica la traducción, a modo de notario? En ese caso, no tiene sentido cobrar por palabra. La traductora ya no produce palabras. Las palabras las produce la máquina. La traductora es una experta que certifica con una posedición, da garantías, avala la fiabilidad. El valor social de ese trabajo no puede calcularse contando por palabra.

Pero no solo está la traducción profesional. Gracias a los traductores automáticos, la traducción es una actividad social extendida al alcance de cualquiera (que tenga un dispositivo, acceso a Internet, etcétera). Hay un mercado lingüístico mundial en el que no todos los idiomas están igualmente representados. En Nigeria hay 203 millones de habitantes y 522 lenguas vivas. El inglés es la lengua oficial, pero el hausa es la lengua más hablada, la hablan 80 millones. Las siguientes son el yoruba, con 50 millones, y el igbo, con más de 30 millones. La presencia de estas lenguas en Internet es prácticamente nula. No hay datos baratos disponibles para entrenar un traductor. En el Reino Unido todas las lenguas insulares, la mayoría de las cuales tienen un origen distinto al del inglés, están oficialmente en peligro de extinción.

Se calcula que solo el 1 % de las traducciones son profesionales. La traducción acerca a las personas, pero ¿al precio de aplanar la comunicación? ¿Se va a simplificar el uso de las lenguas para hacerlas más fácilmente traducibles, es decir, estereotipables, intercambiables…?

¿Qué es una lengua? ¿Para qué sirve? ¿Qué nos proporciona? Todas las distopías negativas imaginan sociedades tiránicas en las que el uso de la lengua está vigilado por un gran hermano formateado por completo, totalmente programado, previsible, sin asumir riesgos —riesgo de error, riesgo de errancia, de duda…—. Relaciones humanas —afectivas, de cooperación, de amistad, amorosas o conflictivas— en las que los riesgos inherentes a la lengua —el malentendido, la ambigüedad, la incomprensión…— desaparecen. ¿Necesitamos una relación con la lengua que resista a todos los intentos de formatearla, programarla, calcularla? ¿Necesitamos mantener abierta la posibilidad de una invención singular, sin importar que no sea traducible?

Los traductores estadísticos y neuronales se entrenan con corpus paralelos de traducciones previamente hechas por humanos. Trabajo

muerto que chupa trabajo vivo. En la medida en que el grueso de las traducciones vayan siendo automáticas, ¿en un futuro cercano se entrenará a las máquinas con corpus engrosados con sus propias traducciones? ¿Entraremos en un proceso recursivo de autorreferencia de las propias máquinas, en un colapso del conocimiento y de las formas del lenguaje que omita, que impida el acceso a lo minoritario, lo singular, lo poco probable?

El advenimiento de los traductores automáticos plantea cuestiones que no solo afectan a lingüistas, traductoras, investigadoras, filósofas, informáticas, matemáticas, ciudadanía global o movimientos anticolonialistas. Por lo visto, las lenguas naturales y las proteínas tienen una estructura similar. Igual que la lengua se puede usar de modo creativo, se están entrenando inteligencias artificiales para tener una especie de ChatGPT que entienda la estructura de las proteínas y genere diseños creativos, proteínas artificiales que no se han dado nunca en la naturaleza, que no han existido en la historia de la evolución. El proyecto ProGPT2 ha usado cincuenta millones de secuencias de proteínas naturales para entrenar un modelo que lo hace.

El conocimiento salta de un ámbito a otro. Cuando uso un traductor automático, quién sabe si estoy colaborando con una farmacéutica. La tecnología está viva. Ningún asunto tecnológico es solo tecnológico.

Conocer

Carta a Adriana

Hola Adriana, ¿cómo estáis?

Me estoy escuchando el pódcast que me pasaste, el de *Hechos reales*, el episodio 2 de «La Torre», que por cierto he visto que en Spotify lo tienen en castellano y en inglés. Es como un *thriller*. ¡Buenísimo!

Son cosas que se imaginan, pero no es lo mismo imaginarlo que escucharlo documentado en un reportaje basado en hechos reales. Madre mía, ¡lo que pasa ahí, dentro de la Torre Agbar! Pobres moderadoras de contenido, todo el rato viendo suicidios, asesinatos, violaciones… Es como si tuviéramos guardaespaldas que nos protegen de todo lo que está ahí pero no queremos ver, aunque, claro, a cambio de su explotación y de quedar hechas polvo.

Y al final del episodio se entiende todo, todos los porqués, todos los entramados que hay detrás de estos empleos que parecen como guais, pero que son como una especie de esclavitud en libertad. Me recuerda a los *delivery*, que es como que en tu tiempo libre vas un ratito en bicicleta y te llevas un dinerillo, pero de eso nada. Algoritmos, dinero, poder, en fin…

Te voy a pasar unos que me estoy escuchando de mecánica cuántica. No te puedo decir que los entienda al cien por cien, ja, ja, ja; pero, incluso sin entenderlos, me parecen lo más. Al final, lo que vienen a decir es que de un sistema físico podemos tener información, muuuucha información. Pero renunciamos a saber lo que es. No vamos a poder saber lo que es, porque para obtener el conocimiento alteramos el sistema. Al experimentar matamos el entorno.

Un electrón, por ejemplo: podemos conocer qué le va a pasar, cómo se va a mover. Podemos tener mucha información sobre sus cambios. Pero, en lo que respecta a saber lo que es, tiramos la toalla. Ya no vamos al ser de las cosas, sino a la información que

tengo de las cosas. Con esa información haré operaciones matemáticas que darán resultados de probabilidades, de probabilidades de que pasen cosas. Pero no tendré la certeza de que algo va a ocurrir. Solo la probabilidad de que ocurra. Dicho en filosofía elegante, dejamos fuera la ontología y aceptamos que solo hay epistemología. Asumimos que el mundo físico no es conocible de forma completa por los humanos. Que está mucho más allá de nuestra comprensión. Aceptamos pulpo como animal de compañía. ¡Uf! ¡Qué cambio! ¡Flipante!, ¿no?

Me recuerda a algunas estrategias de inteligencia artificial, que vienen a decir algo parecido. Nos conformamos con los qués y asumimos que no vamos a llegar a los porqués…

Bueno, si los oyes, ya me dirás.

Y a ver si quedamos para echar una partida de ajedrez, que me encanta que me ganes. Ja, ja, ja.

Besos.

4. Por dentro

Algoritmos

El aprendizaje automático busca construir máquinas que puedan aprender, modificar su comportamiento y autoajustarlo. Así, la máquina tendrá autonomía para predecir resultados orientados a resolver el problema cuando le entren datos nuevos que no haya visto antes. Es un subconjunto de la inteligencia artificial, pues hay otras técnicas para programar inteligencia artificial que no se basan en el aprendizaje automático.

Cuando hay aprendizaje automático, el humano programa el modelo de aprendizaje. El modelo hará autoajustes internos (calculará las w) para maximizar el éxito respecto al objetivo de su aprendizaje: ganar partidas de ajedrez, predecir el tiempo atmosférico, hacer diagnósticos médicos, reconocer imágenes, recomendar contenidos, etcétera.

Para hacer una máquina de aprendizaje automático hay tres estrategias principales: el aprendizaje supervisado, el aprendizaje no supervisado y el aprendizaje por refuerzo. Para cada una de estas estrategias hay disponible una gama amplia de algoritmos.

En síntesis, un algoritmo se entrena con datos y, una vez entrenado, se obtiene un modelo, es decir, una representación del comportamiento o funcionamiento de algo que existe en la realidad (un modelo de cómo se comportan los clientes, un modelo de cuáles son los rasgos que definen a un gato, un modelo de qué tiempo hará mañana…).

La estrategia de aprendizaje supervisado utiliza datos de entrenamiento etiquetados, que incluyen tanto las entradas como la salida correcta. Etiquetar los datos significa verter el conocimiento humano sobre esos datos, anotarlos con palabras clave, categorías o atributos.

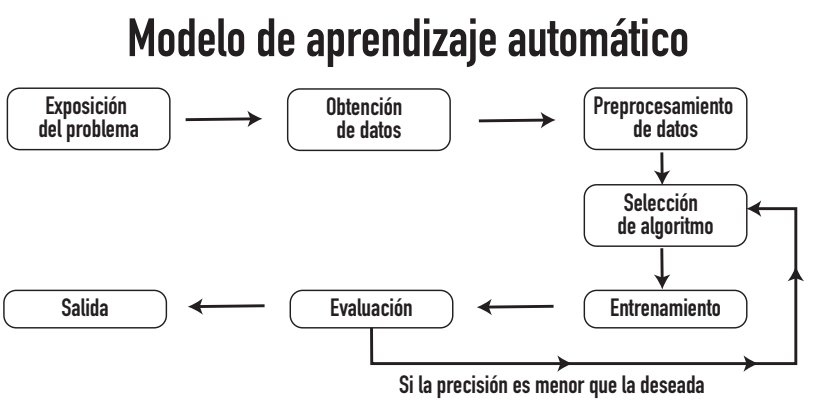


Imagen 17. Proceso de creación de un modelo de aprendizaje automático.

Así, el conocimiento previo sobre los datos se ensambla con estos y ambos (datos y conocimiento previo) son procesados por el algoritmo.

Por ejemplo, etiquetar imágenes puede consistir en adjuntar a las imágenes palabras relevantes sobre su contenido; etiquetar texto, por ejemplo publicaciones en redes sociales, puede consistir en categorizarlo según el tema que trata o su tono emocional; etiquetar audio puede consistir en transcribirlo o identificar a las personas que hablan; etiquetar vídeo puede consistir en identificar objetos, clasificar escenas…

El caso de la profesora que quiere modelizar la nota final considerando un conjunto amplio de factores es un ejemplo de datos etiquetados. Los informes y las notas parciales (la entrada del algoritmo) están etiquetados con la nota final redondeada.

El aprendizaje supervisado se utiliza para clasificar y para pronosticar. Clasificar es lo que hace una niña cuando aprende los colores básicos. Los colores son un continuo, pero se clasifican con unos cuantos nombres concretos: rojo, azul, verde, etcétera.

El caso de la profesora también es una clasificación: clasificar a una persona en el grupo de las que tienen por nota un 5, un 6… Un sistema de reconocimiento de imágenes, por ejemplo qué tipo de especie botánica aparece en una imagen, es una clasificación. Un detector de *spam* en los correos electrónicos también es un clasificador (es *spam* o no es *spam*).

En cambio, pronosticar es estimar resultados futuros. Un ejemplo de pronóstico es el del tiempo meteorológico. La predicción de ventas anuales en una empresa o cuál va a ser el precio máximo por el que se va a alquilar una vivienda también lo son.

El aprendizaje supervisado se puede utilizar con distintos algoritmos o técnicas de cálculo. Uno de ellos es la red neuronal artificial supervisada, pero no es el único. También se puede hacer aprendizaje supervisado con algoritmos que se basan en principios de probabilidad, en métodos estadísticos o en árboles de decisión.

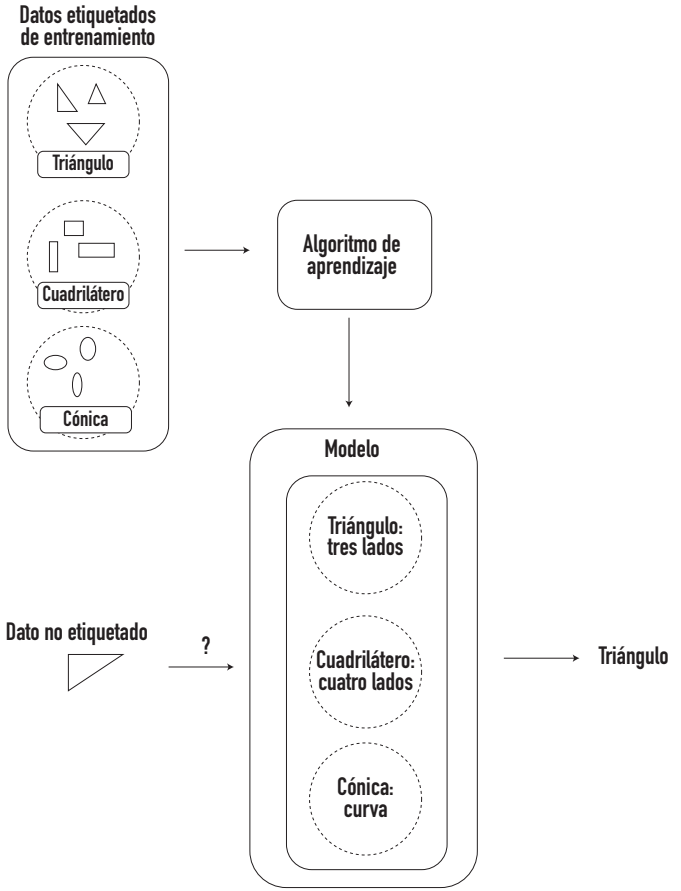


Imagen 18. Entrenamiento y consulta en un algoritmo de aprendizaje supervisado.

El aprendizaje no supervisado se utiliza para descubrir similitudes y diferencias entre los datos, es decir, relaciones o patrones que no siempre son evidentes al ojo humano y que se usarán para hacer agrupaciones o asociaciones (en lenguaje técnico, *clusters*). Estos algoritmos trabajan con datos no etiquetados. Los datos no se pueden etiquetar, porque no hay un conocimiento humano previo o porque no hay capacidad para hacerlo.

Muchas veces, las relaciones o los patrones que están contenidos en los datos no son comprensibles a primera vista, porque son demasiado profundos o demasiado complejos. La idea es buscar sin saber lo que se va a encontrar. Las relaciones o los patrones, si están, emergerán de la estructura subyacente a los datos mismos y la máquina creará por sí misma sus propias etiquetas. El algoritmo, para hacer agrupaciones, calculará una especie de cercanía o de distancia entre los datos y de ese modo formará los grupos.

Para hacer asociaciones, agrupaciones, *clusters*, el algoritmo buscará relaciones. Estas relaciones se pueden usar, por ejemplo, para analizar la cesta de la compra (las personas que compran esto a esta hora del día también compran aquello) y son la base de los sistemas de recomendación de todas las plataformas de contenidos, tipo Netflix o Spotify.

Para el aprendizaje no supervisado se usan redes neuronales, pero también se pueden usar otro tipo de algoritmos, como por ejemplo k-means. La gracia del aprendizaje no supervisado es que sirve para descubrir variables ocultas (en lenguaje técnico, latentes) que no siempre son directamente observables a ojos vista, pero que contienen información esencial.

La diferencia entre hacer agrupaciones con un algoritmo de aprendizaje no supervisado (*clusters*) y hacer clasificaciones con un algoritmo de aprendizaje supervisado es que en las clasificaciones ya se conocen de antemano los grupos que se pueden formar. En el ejemplo de la profesora, serán el grupo de las personas que sacan un cinco, el de las que sacan un seis, un siete… En cambio, con un algoritmo de aprendizaje no supervisado no se sabe cuáles serán los grupos, cuáles las características de agrupación.

Clusterizar es parecido a lo que puede ocurrir en un grupo de personas que no se conocen y van a una convivencia. Las personas, al principio, tienen encuentros aleatorios, se presentan, charlan… y poco a poco se van formando los grupos de afinidad (los *clusters*).

Pero ¿qué es lo que va a definir los grupos? Al principio de la convivencia no se sabe. Quizás se agrupen por aficiones, por ideas, por nacionalidades, por temperamentos, por edades…

El aprendizaje no supervisado se utiliza, por ejemplo, para crear grupos de clientes con comportamientos similares, es decir, perfilarlos. O para detectar anomalías, patrones anómalos, como por ejemplo funcionamientos defectuosos en aparatos o máquinas. O brechas de seguridad en transacciones comerciales…

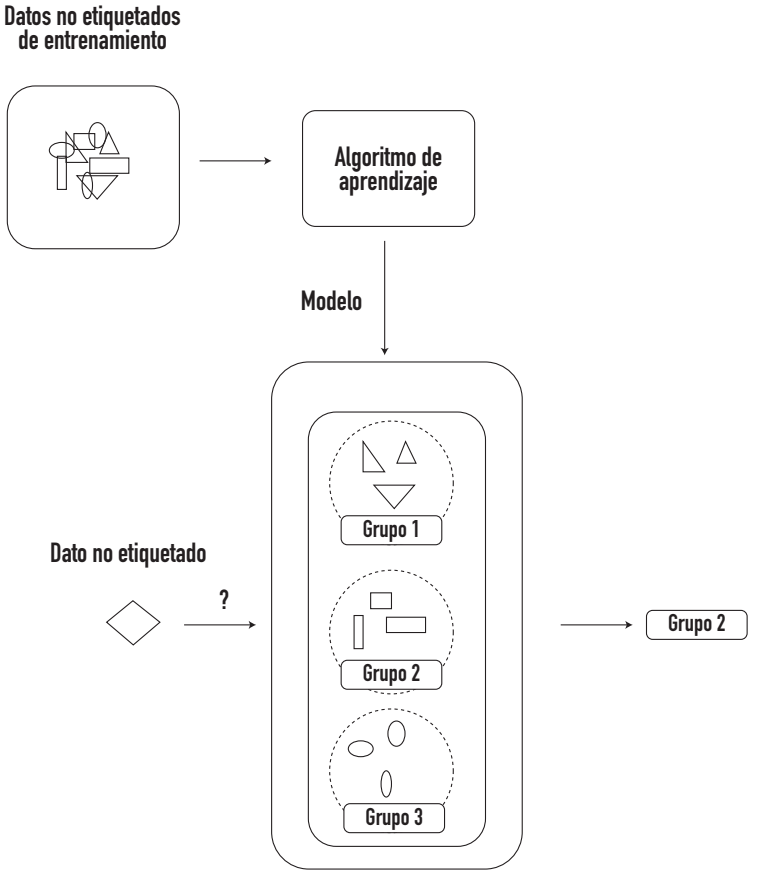


Imagen 19. Entrenamiento y consulta en un algoritmo de aprendizaje no supervisado.

El aprendizaje por refuerzo se usa para resolver problemas en los que hay que tomar decisiones —no una decisión puntual, sino una secuencia más o menos continua de decisiones orientadas a un fin— y es difícil hacer una predicción de cuáles van a ser las mejores. Un ejemplo es la conducción de un vehículo autónomo. Durante el trayecto, el vehículo autónomo está tomando decisiones continuamente, pero es difícil saber cuáles han de ser estas decisiones y no se le puede entrenar con datos, porque la conducción se da en un entorno impredecible y cambiante.

En un algoritmo de aprendizaje por refuerzo hay un agente y un entorno. En lugar de darle datos, se pone al agente a observar el entorno y se le pide que tome una decisión y ejecute una acción. Como consecuencia de esa acción, el entorno cambia. Entonces el entorno informa al agente de la nueva situación (que técnicamente se llama estado) y le da una recompensa o una penalización, según su acción haya sido correcta o incorrecta. Y así hasta que el agente aprende. Es una especie de conductismo.

El agente aprende a generar estrategias por ensayo y error en un ambiente de incertidumbre, dinámico y abierto. A cada acción del agente el entorno cambia de estado, informa al agente de cuál es el nuevo estado y le devuelve el premio o el castigo. El agente va eligiendo patrones de actuación que le den el máximo de recompensas posibles (en lenguaje técnico, va cambiando su política).

Es parecido a aprender a cocinar por ensayo y error, dando a probar platos cocinados a personas que los evalúan. El agente sería la cocinera; la acción sería el cocinar en sí; el entorno serían los comensales; el estado sería cómo están los comensales —aún no han comido, ya han comido esto, ya han comido lo otro…—; el premio o el castigo, la valoración de los comensales; y la política, la manera de cocinar, es decir, el estilo de cocina que va adoptando la cocinera.

Cuando se usa el aprendizaje por refuerzo, no es necesario disponer de un conjunto de datos para alimentar al algoritmo. Los datos se generan con la interacción entre el agente y el entorno. A partir de su interacción con el entorno, el agente va actualizando su política en pro de conseguir el máximo posible de recompensas.

AlphaZero aprendió por refuerzo. En este caso el agente era el programa de ordenador que jugaba al ajedrez, la acción era el movimiento de la pieza elegida, el entorno era el tablero y las reglas del juego, el estado era la posición de las piezas en el tablero en cada momento

concreto de la partida, la recompensa o la penalización era la valoración de la nueva posición que había en el tablero (mejor o peor que antes) y la política era el modo de jugar, la estrategia de juego.

Su red neuronal artificial era del tipo convolucional, que es el que se usa para el reconocimiento de imágenes, porque aprende a abstraer los rasgos fundamentales (por ejemplo, los rasgos de un gato, independientemente de que el gato esté a la derecha o a la izquierda de la imagen, de cara o de espaldas) y puede generalizar, reconociendo gatos que no ha visto nunca. Para el ajedrez, aprendió a abstraer lo importante de las posiciones del tablero ya conocidas y a proyectarlo en las nuevas.

En un mismo contexto puede haber problemas que se solucionen mejor con una estrategia que con otra. En el contexto del ajedrez, valorar numéricamente una posición de piezas en el tablero se podría hacer con aprendizaje supervisado. Los datos etiquetados serían la posición y la valoración que jugadoras experimentadas hacen de esa posición. Si, a partir de un montón de partidas, se quisiera ver qué tipos de aperturas hay, se podría recurrir a un aprendizaje no supervisado. El algoritmo descubriría los tipos de apertura. Si se quisiera una máquina que jugara al ajedrez, se podría hacer con aprendizaje por refuerzo.

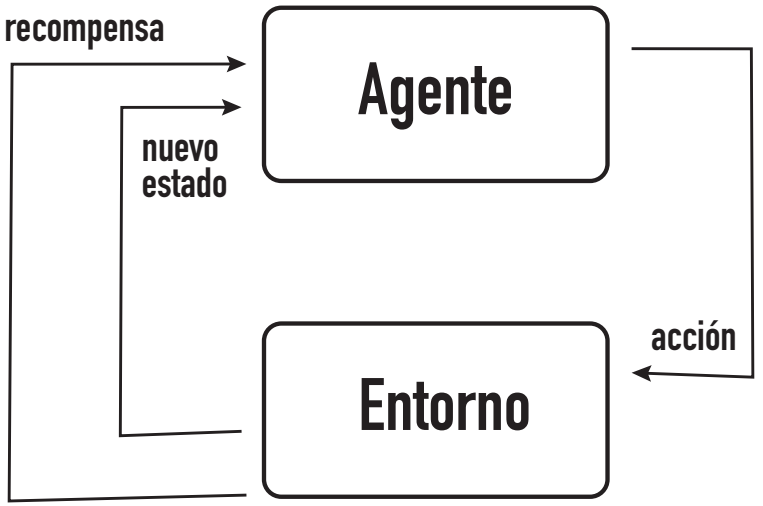


Imagen 20. Funcionamiento de un algoritmo de aprendizaje por refuerzo.

En un contexto escolar, poner notas lo podría hacer un algoritmo de aprendizaje supervisado. Los datos etiquetados serían similares a los de la profesora del ejemplo. Organizar al alumnado en grupos de trabajo que tengan un equilibrio y una diversidad parejos se podría hacer con aprendizaje no supervisado. Y programar una profesora de matemáticas virtual que interactúe con cada alumna y le proponga los contenidos y actividades que necesita muy personalizados, a su ritmo y según su nivel se podría hacer con aprendizaje por refuerzo.

En el contexto de la automoción, un reconocedor de límites de velocidad, que muestre a la conductora en tiempo real cuál es el límite en cada punto de la carretera, se podría hacer con aprendizaje supervisado. Los datos etiquetados serían fotos de señales de límite de velocidad con distintas iluminaciones, enfoques, etcétera, etiquetadas con el número de su velocidad límite. Un detector de anomalías en el motor se podría hacer con aprendizaje no supervisado; y un asistente que aparque el vehículo autónomamente, con aprendizaje por refuerzo.

Para cada problema hay que elegir un modelo y concretarlo hasta el detalle. Si se va a utilizar aprendizaje automático, elegir el algoritmo adecuado tampoco es sencillo. Hay que tomar en consideración muchos aspectos: la calidad de los resultados que se esperan —es decir, cuánto error se puede permitir—; decidir si el modelo tiene que ser explicable o no; ver qué datos hay disponibles o se pueden conseguir y, si son datos sensibles, qué medidas de seguridad hay que tomar; cuánta complejidad se puede manejar, teniendo en cuenta que cuando aumenta la complejidad el modelo es más costoso y menos explicable; si tiene que funcionar en tiempo real, cuál debe ser el tiempo de respuesta; de cuánto presupuesto se dispone; cómo va a evolucionar el modelo y cómo se va a mantener…

Por ejemplo, ¿optamos por un modelo con el que se va a conseguir una precisión del 98 % cuyo entrenamiento va a costar 100.000 € o por otro que tendrá una precisión del 97 % pero cuesta 10.000 €? Hay muchos factores que influyen en la elección del modelo, y a todos ellos hay que añadir la cantidad y calidad de los datos disponibles.

Los algoritmos para construir modelos son públicos y conocidos. Hay mucho conocimiento disponible y mucho software libre. Lo que ya no es tan conocido, lo que no se suele hacer público, salvo cuando hay una apuesta decidida por el software libre, son los detalles, los ajustes con los que se ha precisado el algoritmo: su concreción última.

Datos

Muchas veces la elección del algoritmo está condicionada por la cantidad de datos disponibles y por su calidad: si están limpios, si tienen ruido, si son completos… Dejando a un lado el aprendizaje por refuerzo, programar una máquina de aprendizaje automático requiere datos, muchos datos. Así que la primera pregunta es: ¿hay datos disponibles?

De hecho hay una pregunta previa o cuando menos paralela: ¿hay un modelo matematizable para el problema que la máquina de aprendizaje automático va a resolver? Perseguir la generación de un modelo matemático para un dominio de la realidad, para un trozo de la realidad, supone asumir que esa realidad está regida por leyes y que estas se pueden descubrir.

No es lo mismo hacer una máquina de aprendizaje automático para predecir bajas laborales que hacerla para predecir el tiempo meteorológico. Los modelos de previsión del tiempo son muy complejos, pero los datos con los que se entrena el modelo están muy cercanos a la realidad que representan. Son magnitudes físicas cuantificables muy pegadas al modelo físico. El modelo puede ser muy complejo, pero los datos se desprenden de la observación y son claramente cuantificables. En cambio, respecto a la predicción de bajas laborales el modelo quizás sea más sencillo que el atmosférico, pero entre los datos y la realidad que representan hay mucha más distancia, porque el bienestar o la salud son menos objetivables. Los datos pueden estar teñidos de impresiones subjetivas. No siempre son impepinables.

Un ejemplo de esta subjetividad en el caso de la profesora que pone notas sería el de los informes emitidos por los servicios sociales,

la tutora o la doctora. Dos asistentas sociales, aunque se enfrenten a una misma situación, a la hora de valorarla pueden llegar a diferentes conclusiones, porque, posiblemente, no están viendo la misma situación al cien por cien.

A veces los datos no son suficientes o no son completos. Cuando no son suficientes, se pueden usar datos sintéticos, que se generan, se producen artificialmente. Cuando no son completos, se pueden inferir. Por ejemplo, en un proceso de selección de personal está prohibido requerir datos que revelen el origen étnico de una persona. Sin embargo, ese dato se puede inferir si se combinan otros, como pueden ser el nombre y el código postal del domicilio de la persona candidata. Está claro que no siempre será tan exacto, pero en muchos casos funcionará.

Después de recopilar los datos, hay que prepararlos. Independientemente de que después se vayan a etiquetar o no, los datos hay que prepararlos porque están sucios, contienen ruido, errores, vienen mezclados con información irrelevante… Hay que limpiarlos y normalizarlos.

GPT-3, el modelo de lenguaje que usa ChatGPT, requirió en total 45 terabytes de datos para su entrenamiento. Un terabyte es un millón de millones de bytes: 1.000.000.000.000.

A principios del 2023, después de una investigación, *Time* publicó que OpenAI, la empresa propietaria de ChatGPT, usó trabajadores kenianos subcontratados por la empresa Sama para que revisaran los textos e imágenes con los que se iba a entrenar la máquina con el fin de eliminar relatos horrorosos de abusos sexuales, violencia, torturas, asesinatos, suicidios… Los trabajadores explicaban que esos contenidos se les quedaban grabados en el cerebro después de terminar su jornada y que la imposibilidad de quitárselos de la cabeza era como una tortura. Cobraban entre 1,32 y 2 dólares por hora (en 2024, el salario mínimo interprofesional en Kenia era de unos cien dólares mensuales).

OpenAI, en un comunicado enviado a *Time*, decía: «Nuestra misión es garantizar que la inteligencia artificial general beneficie a toda la humanidad, y trabajamos arduamente para construir sistemas de inteligencia artificial seguros y útiles que limiten el sesgo y el contenido dañino. Clasificar y filtrar [texto e imágenes] dañinos es un paso necesario para minimizar la cantidad de contenido violento y sexual incluido en los datos de entrenamiento y para crear herramientas que puedan detectar contenido dañino».

Detrás de un chat generativo brillante y pulido, listo para mantener conversaciones amables y corteses, hay mucho trabajo humano, explotación y sufrimiento.

Una vez que los datos están preparados, si es preciso se etiquetan. Por ejemplo, si hay que construir una máquina que reconozca imágenes de plátanos y de manzanas, hay que etiquetar todas las imágenes de entrenamiento con la etiqueta «plátano» o «manzana». La cantidad de datos a etiquetar puede oscilar desde miles a millones de millones. Para reclutar personas que realicen este trabajo, están las plataformas de *crowdsourcing*, como por ejemplo Amazon Mechanical Turk.

Mechanical Turk es una plataforma de Amazon que funciona como un mercado de trabajo a destajo: las empresas publican trabajos que otras personas, las *turkers*, pueden realizar de forma remota a cambio de un pago prefijado por la empresa ofertante. Trabajo desregulado bajo demanda. Cientos de miles de personas, en un mercado mundial, compitiendo entre sí para trabajar a precio fijo por unos pocos dólares por hora. La comisión que se lleva Amazon es más o menos el 20 %.

En paralelo a la obtención, limpieza y etiquetado de datos, se va seleccionando cuál algoritmo se utilizará para entrenar el modelo y qué tipo de modelo se desea obtener. Pero no basta con elegir un modelo genérico. Ese modelo se tiene que determinar, se tiene que concretar al detalle. Si el modelo que se elige contiene una red neuronal, hay que precisar qué tipo concreto de red, cuántas capas va a tener, cuántas neuronas va a tener cada capa, etcétera. El modelo de ChatGPT-3 está basado en aprendizaje no supervisado y parametrizado con 175.000 millones de valores (las w). Esa cifra forma parte de la definición del modelo y establecerla no es una ciencia exacta. Es una tarea humana mezcla de matemáticas y programación de ordenadores.

Una de las cosas que se tienen que determinar mediante una decisión humana es la precisión del modelo, es decir, cuánto error es permisible.

En abril de 2024, la revista independiente *+972 Magazine*, dirigida por un grupo de periodistas palestinos e israelíes que, según sus propias palabras, creen en «un periodismo preciso y justo que ponga de relieve a las personas y comunidades que trabajan para oponerse a la ocupación y el *apartheid*, y que muestre perspectivas que a menudo se pasan por alto o se marginan en los relatos dominantes», publicó una investigación sobre Lavender.

Lavender es una inteligencia artificial que estaría utilizando el ejército israelí para identificar objetivos humanos palestinos, presuntos militantes de Hamás o de la Jihad Islámica Palestina, y matarlos. Antes de la operación Espadas de Hierro, para el ejército israelí un objetivo humano era un alto cargo militar palestino, al que podían matar en su casa aunque en la operación también murieran civiles, en muchas ocasiones su familia. Tales objetivos humanos se señalaban con meticulosidad y solo se bombardeaba en sus casas a altos mandos militares para mantener el principio de proporcionalidad acorde al derecho internacional. Lavender era un auxiliar que facilitaba esa identificación.

Pero después del 7 de octubre de 2023 Israel cambió de estrategia y pasó a señalar a todos los operativos del ala militar de Hamás como objetivos humanos, independientemente de su rango o jerarquía militar. Y esto planteó un problema técnico. La cantidad de objetivos humanos a marcar ya no permitía hacerlo humanamente, cotejando pruebas. Había que delegar en una inteligencia artificial automatizada. En las primeras semanas de la guerra, Lavender señaló hasta 37.000 palestinos como presuntos militantes y sus viviendas como posibles objetivos de ataques aéreos. Se trataba de automatizar la identificación de los objetivos.

Según *+972 Magazine*, la autorización para adoptar automáticamente las listas de objetivos de Lavender se produjo unas dos semanas después de iniciada la guerra, después de que el personal de inteligencia comprobara manualmente la precisión de una muestra aleatoria de varios cientos de objetivos seleccionados por el sistema de inteligencia artificial. Cuando esa muestra determinó que los resultados de Lavender habían alcanzado un 90 % de precisión en la identificación de la afiliación de un individuo con Hamás, el ejército autorizó el uso generalizado del sistema. Es decir, se dio por válido un modelo con un error del 10 %. (Sobre el error hablaremos más adelante. Se entiende que de cada cien personas que Lavender señalaba como militantes de Hamás o de la Jihad Islámica Palestina, diez no lo eran. Pero la investigación no informa sobre cuántas personas que sí son militantes no eran señaladas. Dicho de otra manera, no da el dato de a cuántas verdaderas militantes Lavender no identificaba como tales. Quizás ninguna, quizás no se puede saber. El caso es que la investigación no dice nada al respecto).

La máquina Lavender se entrenó con datos de personas pertenecientes al ala militar de Hamás o a la Jihad Islámica Palestina

(información del móvil, conexiones a redes sociales, información del campo de batalla, contactos telefónicos, fotos…). El modelo aprendió a identificar características de esas personas y ahora se aplica a la población en general. Evalúa la presencia de esas características y a partir de ahí la probabilidad de pertenencia, asignando una calificación del 1 al 100. Una persona que presente varias de esas características alcanzará una calificación alta y, por tanto, se convertirá automáticamente en un objetivo potencial. Un automatismo que ahorra tiempo; eso sí, con un error del 10 %.

Más allá de que un 10 % de error en un asunto de vida o muerte es una brutalidad, desde la frialdad de los datos se plantea otra cuestión técnica, o de modelo: qué se entiende por pertenecer. La pertenencia es un límite impreciso. Una persona que no está incluida en la nómina de Hamás, pero ayuda con algún tipo de soporte, ¿es un operativo de Hamás? ¿Lo es alguien que antes estuvo en Hamás, pero que ahora ya no forma parte de esa organización? ¿El teléfono que se está rastreando es verdaderamente un teléfono personal o es costumbre pasarse teléfonos y compartirlos entre unas personas y otras? El listón se puede poner más o menos alto y las características del modelo son muy imprecisas.

Cabe decir que el ejército israelí rechaza rotundamente estas informaciones y niega que esté usando inteligencia artificial para decidir objetivos, alegando que se trata simplemente de herramientas auxiliares que ayudan a los oficiales en el proceso de incriminación.

Sea cierto o no su uso automático, la cuestión es que las fuentes de datos que se van a utilizar, las consideraciones sobre la calidad y la precisión de los datos disponibles, así como el establecimiento de otras características del modelo, como por ejemplo qué se entiende por pertenencia o cuál es el error aceptable, todas ellas son decisiones humanas que se toman en el marco de circunstancias concretas en medio de las cuales la inteligencia artificial es solo un componente generado por humanos y ensamblado en una larga cadena, o red, de la que forman parte muchas decisiones y funcionamientos, que en el caso de Lavender incluirían la de matar al objetivo en su domicilio cuando hay otras personas, hacerlo utilizando bombas baratas que tienen poca precisión, etcétera. Y este ensamblaje no se produce solo cuando hay una guerra. Siempre hay un ensamblaje cuando se arroja al mundo un dispositivo tecnológico.

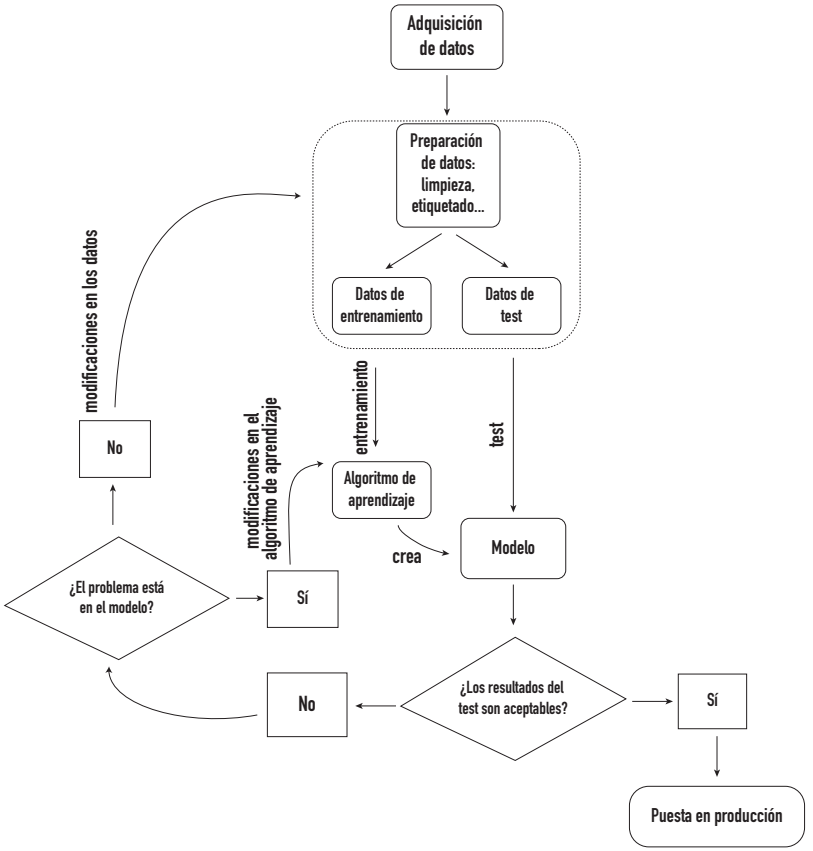


Imagen 21. Proceso de entrenamiento y testeo de una red neuronal artificial supervisada.

En el proceso de entrenar un algoritmo, cuando este se tiene completamente definido y los datos están preparados, estos se agrupan en dos conjuntos: el de entrenamiento y el de test.

En primer lugar, se entrena la máquina con los datos de entrenamiento. Después del entrenamiento se utilizan los datos de test para medir la precisión del modelo. Si los resultados del test no son aceptables, se vuelve a entrenar al modelo con nuevos datos o se hacen modificaciones en el algoritmo de aprendizaje. Si son aceptables, el modelo se da por validado y se puede poner en funcionamiento, en producción.

Cuando se dice que una inteligencia artificial aprende sin intervención humana, a partir de los datos, conviene poner el acento en la cantidad de intervención humana que es necesaria para que se produzca ese aprendizaje automático. Los datos se tienen que identificar, obtener, preparar y dar por buenos. El modelo, el enfoque computacional, se tiene que definir al milímetro. El testeo y la validación también son actividades humanas. Y el ensamblaje de esa máquina en el entorno social donde va a operar como una pieza más en combinación con otras piezas tecnológicas, sociales, políticas o económicas está sujeto a infinidad de decisiones relevantes ajenas a la tecnología, tomadas por personas o grupos humanos. En la producción de inteligencia artificial la cantidad de trabajo humano es enorme: grupos y personas inmersos en entramados de alianzas, conflictos y relaciones de poder, más o menos subalternos entre sí, que condensan en la máquina sus visiones, intereses u objetivos, pero también sus penalidades, sufrimiento y pobreza. Es lo que Mary Gray llama el trabajo fantasma.

Las corporaciones se ufanan de lo rápido que aprenden sus inteligencias artificiales, de lo eficientes que son sus algoritmos. Pero no explican cuánto trabajo humano, en diferentes jerarquías y con diferente reconocimiento social y económico, hay detrás. Los costes medioambientales y sociales no entran en la ecuación.

Se suele pensar que las decisiones humanas relativas a la inteligencia artificial las toman las informáticas. Es cierto que las informáticas pueden tener la responsabilidad en el diseño de un modelo y en su validación, pero la superestructura en la que se enmarca la determinación de qué problema se va a resolver y qué se entiende por haberlo resuelto escapa del dominio de lo puramente algorítmico. Las informáticas no suelen decidir cuánto dinero hay que pagar a las trabajadoras en Kenia ni cuál es el margen de error aceptable para un modelo que decide dónde se tiran las bombas.

Al final de todo este proceso de entrenamiento de una máquina, en el mejor de los casos, se habrá obtenido un modelo muy especializado (si está entrenado para reconocer manzanas, no podrá reconocer peras) que proyecta el pasado en el presente (dice que algo es en función de lo que otras cosas en el pasado fueron) o el pasado en el futuro (predice lo que ocurrirá en función de lo que antes ocurrió), dando una probabilidad de que algo sea cierto dentro de un margen de error determinado.

Pero cuidado, porque se está modelizando una realidad que puede cambiar. Por ejemplo, las meteorólogas están alertando de que ha habido muy buenos modelos para pronosticar el cambio climático. Sin embargo, precisamente porque ahora está aconteciendo ese cambio, los modelos actuales de pronóstico podrían dejar de servir (la proyección del pasado en el futuro ya no predeciría bien) sin que todavía se puedan entrenar nuevos modelos, ya que no hay datos suficientes que den cuenta de cómo son o serán los climas después del cambio. Con lo cual podríamos estar entrando en un periodo de discordancia de los modelos actuales.

Y esto que pasa con el tiempo atmosférico puede pasar con todo, desde la violencia de género hasta el correo *spam*. Y de hecho pasa.

Sesgo

La palabra «sesgo» tiene muchos significados. Hay sesgos cognitivos humanos, como por ejemplo el sesgo de confirmación, que da mucha más relevancia a los datos que confirman una idea previa que a los que la desmienten.

Hay sesgos estadísticos. Por ejemplo, una investigación sociológica sobre una guerra solo puede hacerse contando con las personas supervivientes. Una investigación sobre violencia de género solo puede hacerse contando con mujeres que se autorreconocen como víctimas. Poder contar solo con una parte (sesgada) de las personas involucradas puede introducir sesgos. No siempre se trata de una voluntad de manipulación, sino que a veces solo se dispone de datos parciales, es decir, sesgados. En estadística se entiende que el sesgo es un error sistemático en el que se puede incurrir cuando al hacer muestreos o ensayos se seleccionan o favorecen unos datos frente a otros.

Los movimientos en pro de la justicia social hablan de sesgo algorítmico cuando una inteligencia artificial genera errores (muchos o pocos) sistemáticos y repetidos que perjudican o privilegian a unos grupos sobre otros. Es decir, cuando reproduce relaciones de poder que castigan siempre a las de abajo.

Y en el testeo y la validación de modelos de inteligencia artificial, en lenguaje técnico se habla de sesgo para referirse a un componente, una parte, del error en la precisión.

El hecho de utilizar la misma palabra para indicar, por una parte, una injusticia sistemática (visión social) y, por otra, un componente

del error de precisión del modelo (visión técnica) complica la comprensión de cada uno de estos significados. Por eso, utilizaré sesgo para referirme a lo técnico y discriminación para referirme a la injusticia estructural —lo que los movimientos sociales llaman sesgo algorítmico—, con la intención de diseccionar estos dos significados independientemente y luego ver cómo se relacionan.

Desde el punto de vista técnico, en un modelo de inteligencia artificial se entiende por error la producción de resultados, predicciones, no precisos. El error no tiene un único origen, sino que se puede colar por distintas rendijas: tiene partes, componentes. Por eso se puede descomponer, se puede descuartizar y se puede aislar qué parte del error se ha colado por cuál rendija. Uno de esos componentes es el sesgo. Por error se entiende el error total, en el cual el sesgo es solo una parte, una componente.

Cada componente del error tiene su fórmula matemática, pero intuitivamente podemos decir que hay tres rendijas, tres fuentes de error: el error por defecto, el error por exceso y el error irreductible.

Sobre el error irreductible hablaremos luego, pero, respecto al error por defecto y al error por exceso, tan malo es no llegar como pasarse.

¿Qué es no llegar? No llegar ocurre cuando el modelo, después del entrenamiento, no ha pillado el patrón de relación que hay entre las características relevantes y los resultados correctos.

Por ejemplo, en una tienda de alquiler de bicicletas ven que los días con mal tiempo no alquilan ninguna. Quieren tener una inteligencia artificial que prediga si tal día alquilarán o no bicicletas para decidir si ese día abren o cierran. Supongamos que definen como mal tiempo los días cuya temperatura máxima es menor que 5ºC y la probabilidad de lluvia mayor que 60 %. Está claro que solo con esos dos factores la inteligencia artificial no va a funcionar. En la caracterización de lo que es mal tiempo hay muchos más factores: la velocidad del viento, la humedad, la niebla, la nubosidad…

En este caso, el modelo no alcanza para captar la complejidad de la realidad que quiere representar. Los supuestos son demasiado simplistas. Este error por defecto, por falta de complejidad, se llama sesgo (en inglés, *bias*).

El sesgo se puede deber a que la inteligencia artificial se ha entrenado con pocos datos o a que son demasiado simples, es decir, a que no tienen en cuenta todos los factores influyentes.

También puede deberse a un algoritmo de aprendizaje demasiado simple. Por ejemplo, una red neuronal para clasificar fotografías entre cuarenta colores dominantes tiene que tener más neuronas que una para clasificar fotografías entre las que son en color y las que son en escala de grises, porque en cuarenta colores hay más factores influyentes. El modelo válido para color o escala de grises es demasiado simple para una gama de colores amplia.

Finalmente, este sesgo, o error por defecto, también puede deberse a que el modelo se entrenó muy poco tiempo. Digamos que iba por buen camino, pero no le dio tiempo a terminar de aprender.

¿Qué se puede hacer para reducir el sesgo? Se pueden recoger nuevos datos, añadirles más características, mejorar el algoritmo de aprendizaje o darle más tiempo de entrenamiento. ¡Ah, estupendo! Entonces todo resuelto, ¿no?

Bueno, no tanto, porque también existe el riesgo de pasarse. Pasarse es lo que ocurre cuando el modelo sabe demasiado y, de tanto como sabe, se vuelve tonto. Sabe mucho, pero no de lo que tiene que saber. En este caso hay un error por exceso.

El error por exceso se llama varianza (en inglés, *variance*) y se puede producir cuando los datos de entrenamiento contienen ruido, es decir, características que no son relevantes. Por ejemplo, se entrena un modelo para que reconozca fotografías de perros y gatos. Pero en el entrenamiento la mayoría de las fotos de perros eran fotos de exterior y la mayoría de las fotos de gatos eran fotos de interior. El modelo recibe una información que no necesita. Si es interior o exterior, es irrelevante para distinguir entre un perro y un gato. Es más, es una información que molesta. No es información, es ruido. Como se ha entrenado al modelo con datos que tenían mucho ruido, el modelo ha aprendido a interpretar el ruido como información. No ha extraído el patrón subyacente, porque lo ha hecho demasiado complejo. Está sobreinterpretando la información. Se pasa por exceso, porque tiene en cuenta factores que no vienen al caso (interior o exterior). Se enreda en los detalles. Cuando se le presenten nuevas fotos de perros en interior o gatos en exterior, no los reconocerá bien.

Este error se ha producido porque los datos no eran los adecuados. Tenían mucho ruido. Pero el error de varianza también puede ocurrir cuando los datos de entrenamiento no representan toda la variabilidad de situaciones posibles, cuando la complejidad del modelo es demasiado alta para la realidad que tiene que manejar o cuando se

le ha entrenado demasiado tiempo. En todos estos casos el modelo se pasa de rosca.

Imaginemos que en un intento ideal de llegar al error cero se pone en marcha un proyecto para fotografiar todos los gatos y todos los perros del mundo en distintos escenarios, posiciones, etcétera. Si se entrena el modelo con esos buenísimos y completísimos datos y el modelo tiene la complejidad adecuada, entonces no hay error posible. ¿Cierto?

No tanto. La cuestión es asegurarse de que, mediante el entrenamiento, el modelo capta el patrón subyacente a los datos y podrá funcionar bien a futuro. El modelo tiene que descubrir y extraer las relaciones entre las características de los datos. Tiene que descubrir el patrón, la norma. El problema irresoluble es que en la realidad, en la vida, por más objetividad que se persiga en los datos, siempre habrá datos que se saltarán su propia norma. Siempre hay excepciones. En el mundo real hay perros que parecen gatos y gatos que parecen perros. Así es la vida.

Entonces una de dos: o le pedimos al modelo que memorice o le pedimos que aprenda.

Un camino es darle todas las fotos de todos los gatos y perros del mundo y esperar que se las aprenda de memoria. En ese caso, funcionará muy bien con todos esos datos que ya conoce, pero no funcionará bien con fotos nuevas, porque, si memoriza, tendrá conocimiento sobre un motón de fotos, pero no de las relaciones entre sus características. Se sabrá de memoria un montón de casos sueltos que no responden a ninguna norma. El modelo no ha aprendido a generalizar. No funcionará bien con fotos nuevas. Tiene error por exceso. No es lo que se espera de una inteligencia artificial.

Imaginemos que se quiere enseñar a sumar a una niña y se le ponen siempre las mismas tres sumas. La niña se aprenderá de memoria esas tres sumas y las resolverá con precisión y velocidad, pero en realidad no está aprendiendo a sumar, porque no ha aprendido a generalizar lo que es una suma. Si se le pone una de esas sumas, parecerá que sabe sumar, pero lo que está haciendo es tirar de su memoria. Si se le presenta una suma que no es ninguna de esas tres, cometerá errores. Este aparente, pero falso, aprendizaje, este error por exceso es lo que hay que evitar a toda costa cuando se entrena un modelo.

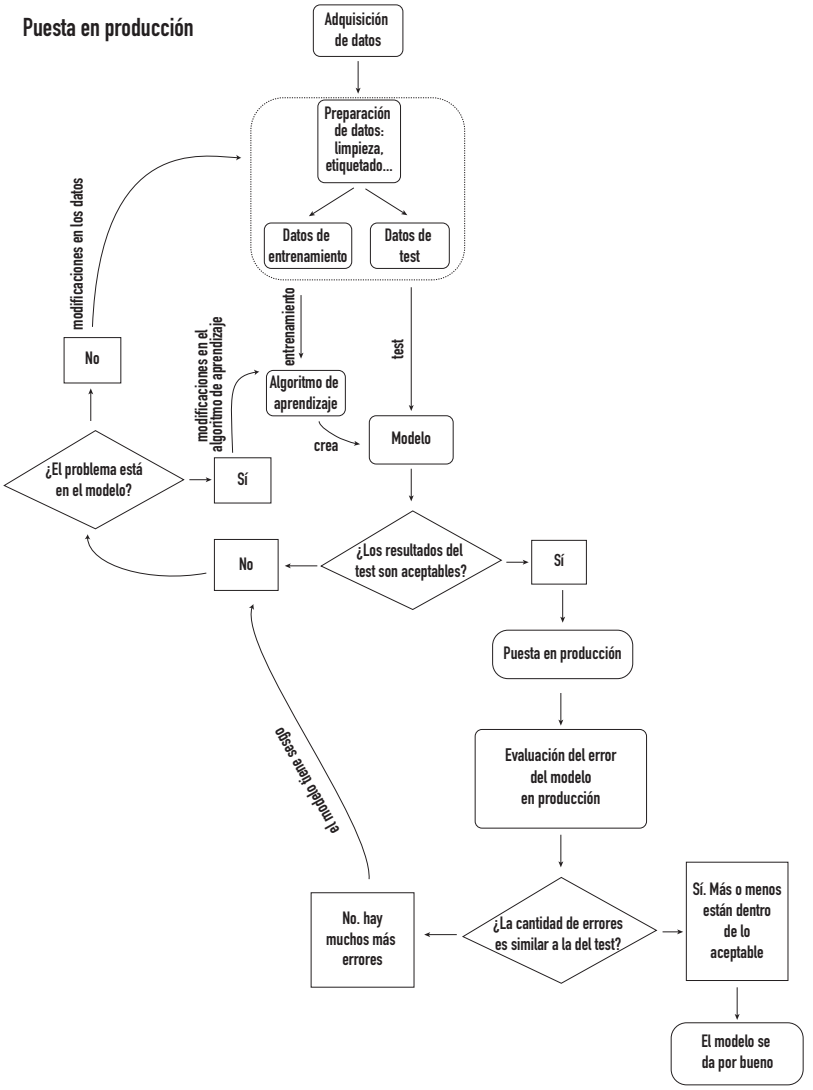


Imagen 22. Una red neuronal en funcionamiento se sigue testando para ver si el error se mantiene dentro del margen aceptable.

Si al modelo le pedimos que aprenda, lo que le estamos pidiendo es que extraiga las relaciones entre las características de los datos y luego, en cierta manera, se olvide de los datos concretos con los que se le ha entrenado y se quede solo con el aprendizaje, con el patrón. De esta manera, funcionará bien con datos nuevos que no ha visto antes. Dicho de otra manera, le pedimos que aprenda a generalizar.

La cuestión es que el modelo, para poder generalizar, tiene que captar la ley de la mayoría. Tiene que quedarse con los perros que parecen perros y los gatos que parecen gatos, y descartar los datos que se salgan de esa ley, de ese patrón. Para generalizar tiene que reducir la complejidad de los datos. El precio de la generalización es sacrificar las excepciones. De alguna manera, el modelo considerará los gatos que parecen perros y los perros que parecen gatos como errores, como ruido que no tiene que tragarse. Al hacerlo, el modelo introducirá algo de ese error por defecto que en lenguaje técnico se llama sesgo.

De hecho, a sabiendas de que esto ocurre así, los modelos que están funcionando, en producción, se siguen testando para saber si el error que se produce, ya con los datos de la vida real, sigue estando dentro del margen de error aceptable. (Ver imagen 22, p. 151).

¿Eso significa que lo más a lo que se puede llegar es a un error bajo, pero no nulo? Bueno, de hecho significa algo mucho más impactante: el error es necesario. Un modelo con error de entrenamiento cero no es un buen modelo. Los datos del mundo real tienen ruido, tienen fluctuaciones. La realidad real no es perfecta, no sigue leyes de exactitud milimétrica. Hay casos que se salen de lo previsible, de la norma. Así son las cosas. En esos casos, el modelo ¡debe! dar error. El modelo no puede atender a la vez a la norma y a las excepciones. No se le puede pedir que se ajuste perfectamente a los datos de entrenamiento y a la vez que aprenda a generalizar. No puede estar en misa y repicando.

El precio que se paga por captar el patrón general, común, a los datos es dejar fuera las excepciones. Pretender que un modelo no tenga error en realidad lo que hará es aumentarlo.

Además de los errores producidos por sesgo y por varianza, hay una tercera componente del error que es el error irreductible (en inglés, *irreducible*), el que no se puede eliminar de ninguna de las maneras. Con el sesgo y la varianza se pueden hacer cosas para reducirlos, para minimizarlos. Pero en la vida real siempre hay elementos fuera de control que se cuelan por la rendija de error irreductible. Este error entra con los datos y es debido a que los datos siempre van a

tener algo de ruido. Nunca vendrán completamente limpios. La realidad no es pura. Los datos juegan al camuflaje.

Un modelo ideal debería ser preciso y consistente, es decir, no equivocarse nunca (tampoco con las excepciones) y poder generalizarlo todo. La cuadratura del círculo. Este ideal es inalcanzable, pero se puede aproximar con un juego de equilibrios. Ajustar un modelo es darle la complejidad justa para manejar la realidad que tiene que modelizar. No darle más ni darle menos. Es entrenarlo el tiempo exacto para que aprenda lo más posible sin pasarse de rosca y con un conjunto de datos del tamaño adecuado, con la máxima representatividad y con el mínimo ruido posibles. Y, con todo eso, con esfuerzo humano y pericia técnica, se podrá minimizar el error. Minimizar, que no eliminar. La tecnología de las redes neuronales, en su estado del arte actual, conlleva asumir que el error es intrínseco. No hay manera de librarse de él.

¡Uf! Y, en este encaje de bolillos, ¿cuál es el error aceptable? El error aceptable depende del contexto. Es una decisión económica, política o social. Lo decide quien tiene poder.

Bueno, es un bajón, pero al fin y al cabo las personas también se equivocan, ¿verdad? Sí, así es. Con algunas matizaciones. Una persona puede ser consciente de que hay algo que no ve claro, de que está dudando, de que necesita más información, de que no puede emitir un juicio. Una inteligencia artificial también, porque hay mecanismos técnicos para que pueda expresar el nivel de confianza con el que ha tomado una decisión. Por ejemplo, puede afirmar que en tal foto hay un gato con el 78 % de confianza. Pero la mucha o poca confianza del modelo respecto a sus propias decisiones siempre se da dentro del marco de su aprendizaje previo, mientras que el marco mental y experiencial en el que se mueve un humano puede ser muchísimo más amplio, más vasto.

Una persona puede explicar qué es lo que tomó en consideración para adoptar una decisión. Una red neuronal no. No es explicable. Y una persona puede ser más sensible a los cambios culturales, puede modificar sus criterios con más agilidad que un algoritmo entrenado. Aunque esto ya es un asunto más de discriminación que de error.

Discriminación

Asumir que los modelos de inteligencia artificial van a hacer predicciones erróneas no es el punto final, sino el punto de partida. A partir de ahí surge un entramado de casuísticas y una amalgama de métricas, con intrincadas y sutiles interrelaciones.

En realidad, las redes neuronales no dan resultados a secas, sino que para cada resultado dan también el nivel de confianza, es decir, la probabilidad que ella misma calcula de que ese resultado sea verdadero. Los niveles de confianza van del 0 (ninguna confianza) al 1 (total confianza).

El problema es que hay que decidir cuál es el umbral de confianza válido. Hay que decidir a partir de qué punto hacia arriba damos por cierto algo y a partir de qué punto hacia abajo lo damos por falso. Hay que trazar una línea que separe los síes de los noes. ¿Dónde ponerla?

Mientras que la minimización de los errores técnicos que se han tratado en el epígrafe anterior suele estar a cargo de las informáticas y las científicas de datos, situar la línea que separa los síes de los noes, es decir, fijar el umbral, es una decisión de negocio o de gobierno. Eso lo decide quien ha encargado la construcción de esa inteligencia artificial (el cliente, la administración…). Dónde poner el umbral es una decisión económica, política…

Para comprender las consecuencias de esa decisión, hay que distinguir entre distintos tipos de acierto y de error. (Ver tabla 1, p. 156).

VERDADERO POSITIVO El modelo predice que es un sí y, efectivamente, en la realidad es un sí. Acierto.	FALSO POSITIVO El modelo predice que es un sí y, por el contrario, en la realidad es un no. Error.
FALSO NEGATIVO El modelo predice que es un no y, por el contrario, en la realidad es un sí. Error.	VERDADERO NEGATIVO El modelo predice que es un no y, efectivamente, en la realidad es un no. Acierto.

Tabla 1. Tipos de acierto y tipos de error.

Sabiendo que habrá errores de falsos positivos y falsos negativos, la cuestión es si ambos errores son igual de costosos o uno es más costoso que otro.

Supongamos que se ha diseñado una inteligencia artificial que hace un diagnóstico sobre la existencia o no de una enfermedad. La probabilidad 0 significa «no hay enfermedad» y la probabilidad 1 significa «sí hay enfermedad».

El modelo da los resultados que se muestran en la imagen 23.

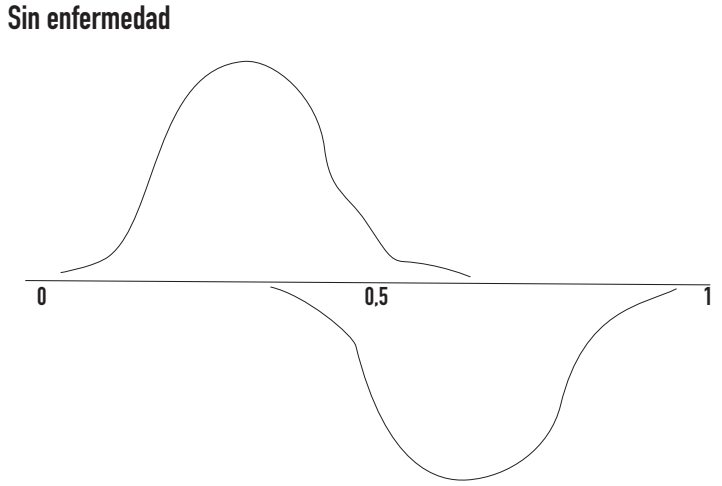


Imagen 23. Resultados de los diagnósticos. Encima de la línea horizontal, los resultados sin enfermedad y debajo los resultados con enfermedad.

Esta inteligencia artificial funciona tipo test, es decir, tiene que dar un sí o un no. Así que el equipo promotor tiene que tomar una decisión. ¿Dónde hay que poner la línea que separe los síes de los noes?

Hay personas que proponen poner el umbral a la mitad, en el 50 %. Así habrá más o menos la misma cantidad de falsos positivos (se diagnostica enfermedad, pero no la hay) que de falsos negativos (no se diagnostica enfermedad, pero sí la hay).

Con el umbral al 50 % quedaría como se muestra en la imagen 24.

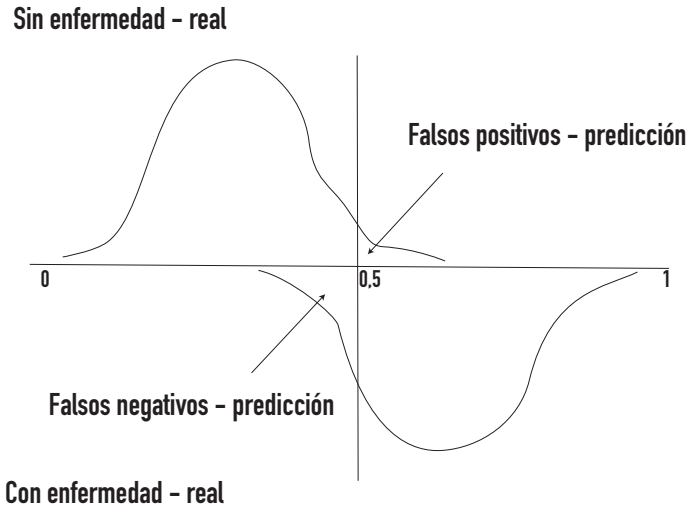


Imagen 24. Umbral con la misma cantidad de falsos positivos que falsos negativos. Encima de la línea horizontal, los resultados sin enfermedad y debajo los resultados con enfermedad.

VERDADERO POSITIVO El modelo predice que hay enfermedad y, efectivamente, la hay. Acierto.	FALSO POSITIVO El modelo predice que hay enfermedad, pero no la hay. Error.
FALSO NEGATIVO El modelo predice que no hay enfermedad, pero sí la hay. Error.	VERDADERO NEGATIVO El modelo predice que no hay enfermedad y, efectivamente, no la hay. Acierto.

Tabla 2. Tipos de acierto y error aplicados a la enfermedad.

Pero no todo el mundo está de acuerdo. Repasan la tabla con los diferentes tipos de acierto y de error y la reescriben como se muestra en la tabla 2.

Hay personas que opinan que no tiene que haber falsos negativos. Deben evitarlos, porque falsos negativos significa personas enfermas que no serán diagnosticadas ni tratadas. De modo que proponen poner el umbral que se muestra en la imagen 25.

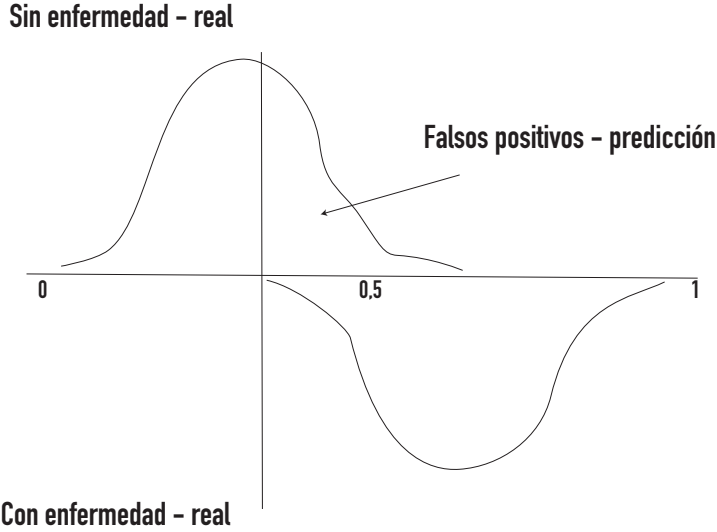


Imagen 25. Umbral sin falsos negativos.

El debate continúa, porque en ese caso habrá muchos falsos positivos, personas diagnosticadas con enfermedad sin tenerla. Se estará medicando a personas que no están enfermas. Y todo medicamento tiene efectos secundarios. Hay que evitar sobremedicar a la población. Solo hay que medicar a las personas verdaderamente enfermas.

Entonces el umbral quedaría como se muestra en la imagen 26, p. 159.

Ya, pero es que entonces, para no sobremedicar, estamos dejando sin tratamiento a personas enfermas.

Lo que hace el equipo promotor con este debate es calibrar las consecuencias de los falsos positivos y los falsos negativos, que no tienen por qué ser simétricas. En este ejemplo, las consecuencias de la calibración son de calado. ¿Qué gravedad tiene esa enfermedad? ¿Cuáles son los efectos secundarios de la medicación? ¿Cuáles son las consecuencias de un diagnóstico positivo? ¿Hay que confinar a esa persona? ¿Qué repercusiones puede tener en la población hacer creer

que hay una epidemia sin tener la seguridad de que la hay? ¿Hay suficientes medicinas para todos los falsos positivos o hay escasez y hay que administrarlas con cuentagotas? El umbral puede colocarse en muchos lugares intermedios.

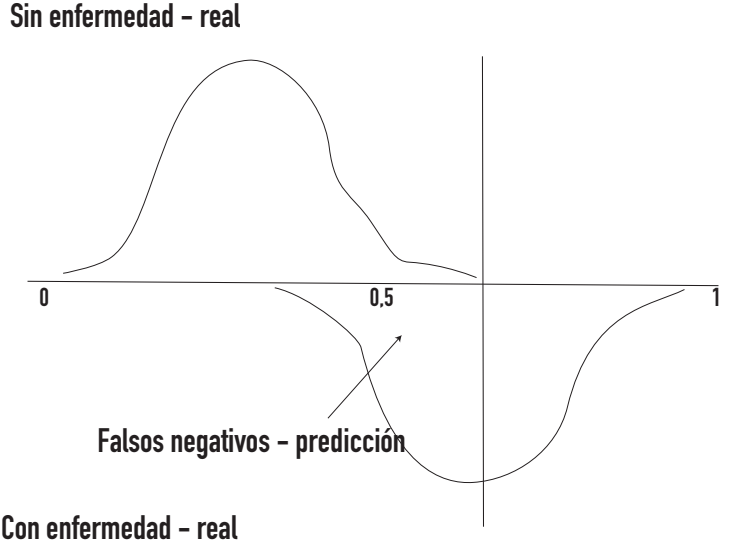


Imagen 26. Umbral sin falsos positivos.

Si se corre hacia la derecha, dará menos falsos positivos y más falsos negativos. Pierde sensibilidad y se hace más exigente, más estricta. Hacia la izquierda, dará más falsos positivos y menos falsos negativos. Gana sensibilidad, pero se hace más laxa, más permisiva.

Hay que situar muy bien el asunto para tomar una decisión que excede lo técnico, porque es una resolución puramente social, política, económica…

Un falso positivo es un resultado incorrecto que se produce cuando la inteligencia artificial identifica como cierto algo que en realidad no es, es decir, indica que algo es cuando en realidad no es. Cuando las consecuencias son punitivas (por ejemplo, negar la libertad condicional), ¿qué es lo aceptable socialmente, lo ético, lo justo? ¿La misma tasa de falsos positivos que de negativos? ¿Negar la libertad solo a las personas que tengan una probabilidad de reincidencia muy alta (evitar falsos positivos), sabiendo que con ese umbral tan a la derecha se dará la libertad a personas que reincidirán?

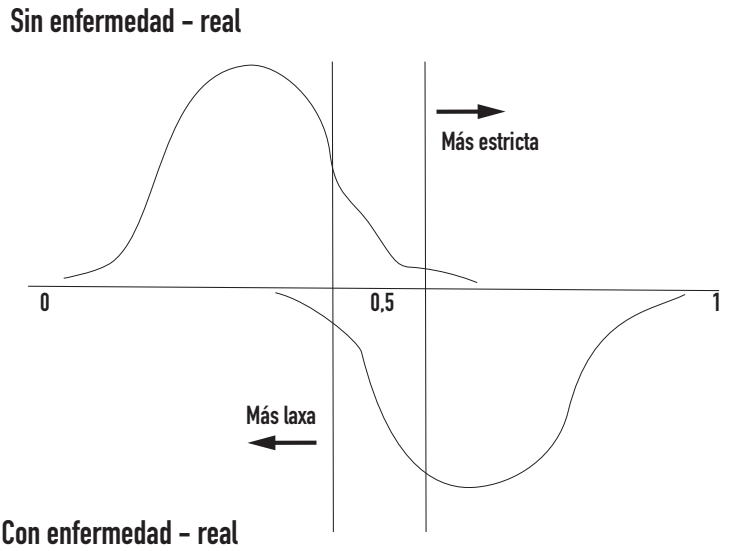


Imagen 27. Consecuencias de mover el umbral hacia la derecha o hacia la izquierda.

Un falso negativo es un resultado incorrecto que ocurre cuando no se identifica algo que debería haber sido detectado, es decir, se indica que algo no es cuando en realidad sí es. Cuando las consecuencias son preventivas (por ejemplo, detectar un cáncer), ¿es deseable la misma tasa de falsos positivos que de falsos negativos? ¿No sería mejor tener más falsos positivos (se le diagnosticó un cáncer que en realidad no lo era) antes que tener algunos falsos negativos (no se le diagnosticó un cáncer que en realidad sí lo era)? ¿No sería mejor mover el umbral hacia la izquierda?

Este tipo de decisiones excede lo meramente técnico y entra en el terreno de la ética, de la equidad y de la justicia.

Trazar la genealogía de una discriminación algorítmica supone identificar por dónde y en qué momento se coló el error y cómo el error se convierte en injusticia. Las usuarias discriminadas por un algoritmo pueden decir, con toda razón, que les da lo mismo el cómo y el cuándo, y que lo que quieren es equidad y justicia. Pero una buena comprensión puede redundar en un crítica de la discriminación más potente.

Supongamos que una empresa quiere una inteligencia artificial que funcione como asistente para la selección de personal para un puesto dado. Buscan los currículums de todas las personas que, a lo largo de la historia de la empresa, han sido seleccionadas para puestos similares a ese y ven que tienen 500 currículums de hombres y 20 de mujeres. ¿Por qué solo 20 mujeres? ¿Porque no hay mujeres que quieran ese puesto? ¿Porque la empresa es machista y prefiere contratar a hombres?

Además, también tienen 2.000 currículums de hombres y 40 de mujeres que fueron rechazados.

Entrenan la inteligencia artificial con los currículums aceptados y los rechazados. Lógicamente, el modelo aprenderá más sobre hombres que sobre mujeres. Hay un grupo (las mujeres) que está subrepresentado. Pero el paso del error a la injusticia no depende solo de esa inteligencia artificial aislada, sino de cómo es el entorno social en el que funciona.

Si para el nuevo puesto solo se presentan hombres, la inteligencia artificial funcionará de maravilla. Los seleccionados serán buenos candidatos. Si se presenta una mujer con un currículum muy brillante, también será seleccionada. La inteligencia artificial tiene precisión. Cuando dice que una candidatura es buena, es porque es buena. No da falsos positivos.

Pero ¿qué pasa si se presentan un hombre con un currículum justito y una mujer con un currículum justito? La inteligencia artificial no puede dudar. Siempre va a dar un resultado. En caso de empate, para minimizar el error, elegirá al hombre. Digamos que eligiendo al hombre arriesga menos, porque eso es lo que ha hecho históricamente la empresa. Por no arriesgar, perpetúa estereotipos. Hay un bucle sutil que se retroalimenta.

¿Y qué pasa si solo se presentan mujeres? Probablemente no elegirá bien, porque no ha aprendido mucho sobre cómo hacerlo. No elegirá con precisión, porque ha habido un cambio social y la inteligencia artificial se adapta mal a los cambios culturales, sociales. Que la máquina selectora de currículums genere discriminación no depende solo de la máquina en sí, sino también del contexto social donde se la pone a funcionar. La discriminación surge en la interacción.

Imaginemos que las mujeres candidatas protestan ante la empresa, argumentando que van a ser discriminadas por el algoritmo. La empresa va al equipo informático y les dice: hay personas que están

protestando. Dadme las métricas del error, que queremos ver qué precisión está dando del modelo. Las métricas dan una precisión muy alta. De todos los positivos que da, el 99 % son verdaderos positivos.

La empresa muestra este dato a las candidatas y les dice: «La inteligencia artificial es muy precisa. Es objetiva. No discrimina». Pero las mujeres no quedan satisfechas. Entienden bien lo que significa esa métrica y pueden contraargumentar. Cuando los grupos no están igualmente representados, esa medida de precisión no sirve, porque lo que hay que ver es si los errores, aunque sean pocos, tienen un patrón que discrimina siempre al mismo grupo. Ese error del 1 % ¿es siempre en currículums de mujeres? La discriminación no viene de que haya muchos errores aleatorios, no viene de que la inteligencia artificial funcione como una escopeta de feria. Viene de que los errores, pocos o muchos, tengan un patrón repetitivo y sistemático que perjudica a un grupo y beneficia a otro.

Por el contrario, imaginemos que la empresa es consciente de que hay un grupo subrepresentado y de que esta subrepresentación puede derivar en que los currículums de las mujeres sean falsos negativos: currículums buenos que la inteligencia artificial deja escapar. Imaginemos que tiene conciencia social y quiere que las contrataciones sean equitativas. ¿Qué puede hacer?

No puede entrenar a la máquina con más currículums de mujeres, porque no los tiene y no se los puede inventar. Podría poner en marcha un procedimiento de doble *check*. Un grupo de personas evalúan todos los currículums presentados y cotejan su evaluación con la que ha hecho la máquina. Comparan los positivos de la máquina con los suyos, y lo mismo con los negativos. Miran las discrepancias y buscan patrones en los errores. Quieren ver si los errores son sistemáticos, si responden a algún factor discriminador. Después de esa investigación, si detectan patrones en los errores, se podría modificar el modelo, ajustarlo con la intención de romper esos patrones de error. Si los errores se pueden ver y hay voluntad, la discriminación puede disminuir e incluso, quizás, desaparecer.

En un entorno ideal, esto se podría hacer. En la realidad es difícil que ocurra, porque la empresa encarga la inteligencia artificial para ahorrar tiempo y dinero en la selección de personal, por lo que es muy poco probable que monte un sistema de doble *check*, que duplicará ese tiempo y ese dinero.

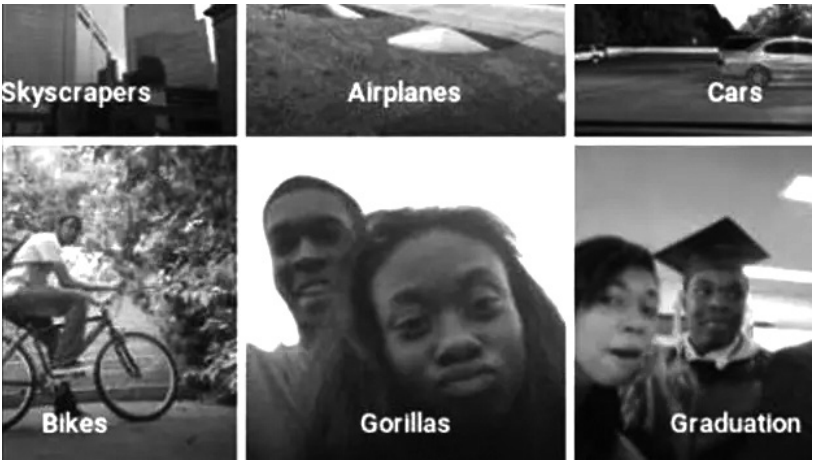
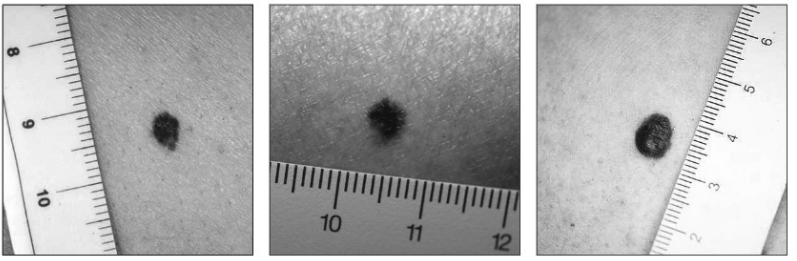


Imagen 28. La inteligencia artificial de Google Photos etiquetó a estas dos personas como «gorilas». (Fuente: https://www.bbc.com/mundo/noticias/2015/07/150702_tecnologia_google_perdon_confundir_afroamericanos_gorilas_lv).



(a) Clark nevus (b) Melanoma in situ (c) Melanoma 0.76–1.5 mm

Figure 1: "Naked eye" images of melanocytic lesions [3].

Imagen 29. Una inteligencia artificial cuyo objetivo es diagnosticar melanomas que haya sido entrenada con fotografías en las que en los casos positivos aparece una regla puede considerar que la regla en sí es un dato relevante, cuando en realidad no lo es. (Fuente: https://onlinelibrary.wiley.com/doi/epdf/10.1155/2018/5767360).

Supongamos el mismo caso: una inteligencia artificial para selección de personal puesta en marcha en un hospital para cubrir puestos de enfermería. Pasaría algo muy parecido, pero al revés. En este caso, ¿la sociedad considerará que hay una discriminación? Que haya o no

haya discriminaciones no es algo que tengan que decir solo las científicas de datos. Es una cuestión social, de valores. Aunque comprender cómo funciona todo esto y establecer alianzas con las científicas puede ser muy potente.

Los datos pueden ser fuente de discriminación. Cuando son etiquetados con trabajo mal pagado, es presumible que contengan errores. También cuando están desequilibrados y no representan por igual a todos los grupos étnicos o sociales. En julio de 2015 Google tuvo que pedir disculpas porque Google Photos, una app que etiquetaba automáticamente con inteligencia artificial las fotos que subían los usuarios, había etiquetado a una pareja de personas negras como «gorilas» (ver imagen 28, p. 163).

A veces los datos están mezclados con características irrelevantes. Por ejemplo, una inteligencia artificial para identificar melanomas entrenada con imágenes en las que en los casos positivos las fotografías incluyen una regla de medir puede considerar que la regla de medir es relevante, cuando en realidad no lo es. Puede privilegiar las fotografías con regla sobre las que no la tienen (ver imagen 29, p. 163).

En estos dos casos, la elección de los datos ha sido mala. El *dataset* se ha diseñado mal. Los datos no representan bien lo que se quiere modelar. La inteligencia artificial resultante no tendrá precisión. En cuanto se ponga a funcionar, los errores (en el caso del melanoma) y la discriminación (en el caso de los gorilas) saltarán a la vista.

En el año 2018 el gobernador de Salta (Argentina), Juan Manuel Urtubey, anunció la implementación de un plan piloto que, con un programa de inteligencia artificial creado por Microsoft, predeciría los embarazos en chicas adolescentes durante los próximos cinco años y aplicaría políticas públicas para prevenirlos. Este anuncio se daba en plena lucha feminista por la despenalización del aborto.

Según una investigación de *Wired*, Microsoft utilizó una base de datos de más de 12.000 mujeres entre 10 y 19 años para crear el modelo predictivo a partir de la edad, el barrio, la etnia, el país de origen, la presencia o ausencia de discapacidad, si se dispone de agua caliente en el baño y si el cabeza de familia ha abandonado los estudios. El modelo no incluía ningún dato sobre el uso de métodos anticonceptivos ni educación sexual. Con base en la predicción, agentes territoriales visitaron las casas de las jóvenes identificadas por el modelo, pasaron cuestionarios, tomaron fotografías y registraron ubicaciones GPS, sin que hubiera ninguna transparencia sobre qué pasó después con esos datos personales.

Grupos de feministas y de periodistas activaron la organización comunitaria para hacer lo que vendría a ser una auditoría popular de un sistema de inteligencia artificial. En sus propias palabras, «la idea de que los algoritmos pueden predecir el embarazo adolescente antes de que ocurra es la excusa perfecta para que los activistas antimujeres y anti derechos sexuales y reproductivos declaren innecesarias las leyes sobre el aborto». Después se supo que el doctor Abel Albino, que se oponía abiertamente al derecho al aborto, estaba detrás de este proyecto, en alianza con Microsoft.

Según esta auditoría popular, el modelo «confunde las variables socioeconómicas para hacer parecer que la niña o la mujer es la única culpable de su situación. Carece totalmente de cualquier consideración del contexto. Este sistema de inteligencia artificial es un ejemplo más de la violación de los derechos de las mujeres por parte del Estado. Imaginen lo difícil que sería negarse a participar en esta vigilancia». Familias pobres que dependen del organismo patrocinador del programa, el Ministerio de Primera Infancia, para servicios como vacunas y leche gratuita no pueden negarse a participar en la encuesta.

Además de todos los prejuicios patriarcales —como considerar que la joven es la única responsable del embarazo o que la educación sexual comunitaria puede sustituirse por programas de control social de la población pobre—, además de cebarse en la pobreza —pues los embarazos adolescentes son mucho más fáciles de ocultar en familias ricas que en familias pobres— y además de todas las consideraciones sobre el derecho a la privacidad y el tratamiento de los datos personales, esta inteligencia artificial tenía un grave error metodológico: los datos con los que se entrenó la máquina estaban tomados en el momento en el que las adolescentes quedaban embarazadas, es decir, en el momento del embarazo. Esos datos no describen lo que aconteció antes, cuáles eran las circunstancias de la joven cinco años antes del embarazo. Por eso, de ellos no se puede extraer un patrón predictivo a cinco años. Para hacerlo, en el mejor de los casos, habría que haber tomado los datos de esa joven que ahora está embarazada hace cinco años.

En este caso, los datos solo capturan las características influyentes parcialmente: son los datos del momento del embarazo y no los datos de cinco años antes. Pero, además, el peso del patriarcado se transfiere al modelo, pues solo se tiene en cuenta datos de la joven, asumiendo que ella es la única responsable. Una inteligencia artificial de este tipo es prácticamente imposible de evaluar, porque interviene en la

situación social que modela. Al intervenir en la vida de las jóvenes, si el modelo da que una joven quedará embarazada y no queda, no se puede saber si es por la aplicación de políticas públicas de control social o por cualquier otra circunstancia, como el entorno de la joven, sus propias capacidades personales… El modelo cubre de una manera muy pobre la realidad que quiere modelizar. Los falsos positivos no se pueden rastrear. Los falsos negativos sí. En 2022, para *Wired* no estaba claro si el programa había sido definitivamente suspendido.

En el año 2012 el Departamento de Instrucción Pública (DPI) del estado de Wisconsin (Estados Unidos) puso en marcha el Sistema de Alerta Temprana de Abandono Escolar (DEWS), utilizado para predecir futuras deserciones escolares en la enseñanza pública secundaria. El sistema puntuaba a cada estudiante con un número —de 0 a 100— que representaba la probabilidad de que se graduase en los cuatro años posteriores a su ingreso en la escuela secundaria. Esta puntuación la calculaba una inteligencia artificial que había sido entrenada con los datos históricos de la frecuencia de la graduación de estudiantes en cursos anteriores. Para puntuar, la inteligencia artificial DEWS tenía en cuenta la asistencia, las expulsiones, la cantidad de escuelas anteriores, la puntuación en los exámenes de inglés y matemáticas, y datos demográficos como la pobreza, la etnia o el género.

Dos veces al año, las escuelas públicas de Wisconsin recibían una lista de sus estudiantes junto con la predicción de riesgo de abandono de los estudios codificada con colores: verde para el bajo (puntuación mayor que 78,5), amarillo para el moderado (puntuación igual a 78,5) y rojo para el alto (puntuación menor que 78,5). DEWS se presentó como una herramienta contra la brecha de graduación. Los estudiantes blancos se graduaban con más frecuencia que los hispanos y los negros. El objetivo era que los educadores dispusieran de datos predictivos con la suficiente antelación para que pudieran prestar más apoyo a esos estudiantes.

Tras una década de implantación, el medio de comunicación sin ánimo de lucro *The Markup*, que investiga cómo utilizan las instituciones influyentes la tecnología para cambiar la sociedad, requirió al DPI los registros públicos sobre los datos relativos a las graduaciones y los analizó. Su conclusión fue que el sistema estaba fracasando en su propósito (mejorar las tasas de graduación de los estudiantes calificados como de alto riesgo) e influía negativamente en la forma en que los educadores percibían a los estudiantes, especialmente a los negros, porque los estigmatizaba. *The Markup* publicó esos datos en GitHub.

La tasa de falsos positivos de esa inteligencia artificial (con qué frecuencia un estudiante que predijo que no se graduaría a tiempo realmente se graduó a tiempo) fue un 42 % más alta para los estudiantes negros que para los blancos y un 18 % más alta para los hispanos que para los blancos, según los datos proporcionados por el mismo DPI, que no fueron totales.

Cuando *The Markup* se dirigió al DPI para cuestionar si la inteligencia artificial DEWS tenía un sesgo racista, el departamento respondió: «No, el análisis de datos no es racista. Son matemáticas, que reflejan nuestros sistemas. La realidad es que vivimos en una sociedad supremacista blanca y el sistema educativo es sistémicamente racista. Este es el motivo por el que el DPI necesita herramientas como DEWS y por el que estamos comprometidos con la equidad educativa».

The Markup entrevistó a estudiantes y educadores en 80 de los más de 400 distritos del estado para averiguar cómo se estaban usando las predicciones, y concluyó que la falta de precisión y esa alta tasa de falsos positivos generaba estigmatización en las personas etiquetadas como de alto riesgo, mientras que las educadoras no recibían ninguna explicación sobre cómo calcula DEWS sus predicciones ni cómo traducir una etiqueta de alto riesgo en una intervención educativa adecuada. La inteligencia artificial identificaba personas en riesgo y les decía a las educadoras: «Ahí tienen un problema. Resuélvanlo». ¡Como si el problema y la solución fueran personales y no estructurales! Las estudiantes identificadas sufrían, se sentían señaladas, predestinadas al fracaso, y además las profesoras no disponían de recursos para mejorar su paso por el sistema educativo.

Esta visión de que el fracaso escolar es un asunto personal estaba incrustada en el modelo, que además tomaba en cuenta la pobreza, la etnia o el género, factores sobre los que los educadores, a pesar de su implicación vocacional, no tienen mucha capacidad de intervención. No se necesita un modelo de puntuación personal para saber que las personas negras e hispanas tienen menor probabilidad de completar sus estudios que las blancas. Las críticas a DEWS apuntaban al modelo y se basaban en que las transformaciones para que haya justicia social deben ser estructurales, no personales, y que la equidad nunca se va a alcanzar en un sistema educativo con permanentes recortes presupuestarios.

En las entrevistas que realizó *The Markup*, los estudiantes manifestaron cómo se sentían cuando los etiquetaban como personas en

riesgo. Algunos de ellos explicaron cómo durante la pandemia de COVID-19, cuando las clases pasaron a ser virtuales, se sintieron aliviados y relajados. Podían respirar y sus notas empezaron a subir. «No tenía la presión social, por ejemplo los profesores alrededor de mí o la administración alrededor de mí». Para este estudiante, tener la pantalla del ordenador entre él y la escuela era una liberación.

Desde el DPI se reconoció que el modelo estaba parametrizado para aceptar hasta veinticinco falsos positivos por cada identificación positiva verdadera. Sin embargo, en diez años tantos falsos positivos no habían reducido la brecha de graduación, por lo que había que preguntarse cómo se estaba ensamblando esa inteligencia artificial en las condiciones concretas de ese sistema social educativo en el que se inscribía. DEWS era una especie de vocero que decía: «Sé que algo va a ocurrir, pero no hay ningún medio para que no ocurra». Es como si no sirviera para nada, pero en realidad es peor que eso, porque las personas señaladas sufrían el estigma sin obtener nada a cambio. ¿Un fracaso de la inteligencia artificial como tecnología predictiva? ¿Un fracaso del modelo concreto de DEWS? ¿Un fracaso del sistema educativo, que no fue dotado con suficientes recursos o no sabía cómo intervenir respecto a las alertas tempranas? ¿Un fracaso estructural de una sociedad injusta clasista, machista y racista? Desde 2023 el DPI está evaluando el futuro de este y otros sistemas de alerta temprana.

Imaginemos que en un paso fronterizo se pone en marcha una inteligencia artificial que, según el rostro de la persona, predice si en su maleta lleva tabaco de contrabando o no. El modelo se ha entrenado con fotografías que ha suministrado la policía. Supongamos que, con los datos recopilados hasta el momento, la mayoría de personas que llevaban tabaco de contrabando en su maleta tenían la piel oscura. Supongamos, también, que el objetivo es que nadie meta una maleta con tabaco de contrabando.

En estas condiciones, el sistema hará sonar la alarma con mucha más frecuencia con personas de piel oscura que clara. Imaginemos que el patrón social no cambia, no hay cambios sociales. Anteriormente, las personas de piel oscura daban verdaderos positivos. Si nada cambia, la inteligencia artificial detectará esos verdaderos positivos. Se abrirá la maleta y se verá que hay tabaco de contrabando. ¡Estupendo! ¡Un éxito! Una inteligencia artificial que funciona de fábula.

Uhm…, pero el objetivo no es que las personas con piel oscura no metan maletas con tabaco de contrabando. El objetivo es que

nadie las meta. Con esta inteligencia artificial no sabemos lo que está pasando con las personas de piel clara. ¿Qué está pasando con los negativos? Si no se abren las maletas de las personas con piel clara, no se puede saber. Para analizar los falsos negativos habría que abrir todas las maletas.

¿Se pueden abrir todas la maletas en un paso fronterizo? Uhm…, depende de la política que se quiera aplicar, pero desde luego que se crearían problemas de logística. ¿Se pueden abrir todas las maletas en franjas de tiempo aleatorias? Uhm…, quizás. ¿Sería más eficiente un sistema aleatorio que abra maletas a voleo, sin tener en cuenta nada más que el azar? Uhm…, depende del contexto social, de las características reales de las personas que meten tabaco de contrabando en sus maletas y de lo que se entienda por eficacia y por discriminación.

El sistema tiene mucha precisión. En el 99 % de las maletas que se abren hay tabaco de contrabando. Es un sistema eficaz. Pero no se sabe cuánto tabaco de contrabando entra sin ser detectado en maletas de personas de piel clara. El 1 % de las personas de piel oscura cuya maleta se abre puede sentir que está sufriendo discriminación. Imaginemos que a raíz de las denuncias por discriminación durante una semana se abren todas las maletas. Se comprueba que el 90 % de las maletas con tabaco de contrabando pertenecía a personas con piel oscura. ¿Tiene sentido, entonces, abrir las maletas mitad y mitad? Se abrirán las maletas de muchas personas de piel clara que no llevan tabaco. Muchos falsos positivos es perder el tiempo, es perder eficacia. ¿Qué hacer en estos casos? Y, sobre todo, ¿quién lo decide?

Al final de este recorrido por los errores y las discriminaciones hay un caso que se lleva la palma. Es cuando el dueño de un algoritmo decide, por sus barbas, que una red social difunda más contenidos de derechas que de izquierdas. Ahí ya no hay métricas que valgan.

De la existencia de la discriminación algorítmica, considerada una forma de injusticia estructural, surge la reivindicación de justicia algorítmica, también llamada justicia de datos, defendida por las organizaciones pro derechos humanos y también por investigadoras de la ciencia de datos y de la inteligencia artificial.

La buena noticia es que la discriminación se puede medir, aunque no hay un estándar sobre cómo hacerlo.

Aequitas es un conjunto de herramientas para auditar y evaluar la justicia algorítmica de inteligencias artificiales en entornos de

clasificación binaria. Está licenciada como software libre y puesta a disposición de científicas de datos, investigadoras de aprendizaje automático y responsables de políticas públicas. Ha sido desarrollada por el Center for Data Science and Public Policy, de la Universidad de Chicago, y su código está disponible en GitHub.

Aequitas surge de la preocupación por el riesgo de inequidad en las inteligencias artificiales, que puede afectar injustamente a personas por motivos de raza, sexo, religión o pobreza, entre otros. A pesar de la concienciación creciente sobre este riesgo, no hay un consenso sobre cuál sería una definición de imparcialidad ni qué métricas deberían utilizarse, así que este conjunto de herramientas permite a las usuarias probar modelos con varias métricas de error e imparcialidad en relación con múltiples subgrupos de población.

Desde la justicia algorítmica se pone énfasis en la transparencia: los datos con los que se entrenan las inteligencias artificiales que impactan en la justicia social tienen que ser públicos o, por lo menos, auditables. Desde un punto de vista de la comprensión del error, también sería una medida de transparencia el acceso a las métricas de evaluación con las que se está midiendo la calidad del modelo, pues esas métricas pueden revelar la letra pequeña.

En resumen, a la pregunta sobre si se pueden hacer inteligencias artificiales sin error, la respuesta es no. A la pregunta sobre si se puede hacer que no discriminen, la respuesta, a mi entender, está abierta, pues hacen falta más recorrido y más alianzas tecnosociales. El error es una medida aritmética y la discriminación es un enjuiciamiento social.

La crítica sobre la injusticia algorítmica debate sobre si es técnicamente posible y exigible implementar inteligencias artificiales neutrales o bien el peso de la equidad no debe recaer sobre las máquinas, sino sobre cómo las predicciones de esas máquinas se insertan, se integran, en lo social. Es decir, se debate sobre si hay que exigir justicia algorítmica a las inteligencia artificiales o bien la justicia, a secas, debe ser exigida al sistema social en su conjunto, considerando el resultado de todas las interactuaciones entre elementos maquínicos y humanos, personales y estructurales, de solidaridad y de poder.

Vivir

Carta a Raquel

Querida Raquel, ¿cómo estáis?

Todavía tengo por leer todos los mails con las sugerencias que me has enviado. Millones de gracias. Y muy buena la idea de las inteligencias artificiales que ya hay ahora en los coches normales y corrientes, los que van por nuestras carreteras, y no digamos en los de conducción autónoma. Son buenos ejemplos para entender que en la vida no todo es ChatGPT. Ja, ja, ja.

Sobre eso que dices de que no estoy poniendo nombres, es aposta. Si empiezo a poner nombres, sale una historia de hombres. Seguro que hubo mujeres, que hay mujeres, pero no tengo tanta capacidad de investigación, así que he cortado por lo sano y ale, ¡sin nombres! Si alguien quiere los nombres, que los busque en Internet, que los encontrará a la primera ;-) Menos en la parte de filosofía, que ahí sí los voy a poner, porque me quedaba raro explicar las ideas sin decir quién las manifestó.

No me va a dar tiempo de mirar todas las series y películas que me has dicho. No doy para tanto y no terminaría nunca. He buscado pelis que traten el tema de la singularidad tecnológica y no me queda claro que *2001: Una odisea del espacio*, o *Terminator*, o *Matrix*, o *Her*, o esas que suelen salir en las listas lo sean. Pero, claro, es que primero tendría que tener claro qué es una singularidad tecnológica. Ja, ja, ja. Creo que una singularidad tecnológica se producirá cuando haya una explosión de inteligencia artificial y las máquinas, ellas solas, empiecen a crear otras máquinas cada vez más inteligentes. No veas el pedazo de página que tiene la «singularidad tecnológica» en la Wikipedia. Hay que leérselo.

He investigado un poco por qué a esa futura hipotética superinteligencia se le llama singularidad. Y creo que viene de las matemáticas. En mates una singularidad es cuando las reglas no se cumplen. Mejor dicho, cuando las reglas se cumplen en general, pero hay un puntito muy chiquitito en el que no se cumplen. Es como que una cosa iba por

buen camino, pero en un momento dado da un salto y se sale de madre: se va al infinito. Es como que las leyes se rompen en un punto y de repente ocurre una cosa que no tiene continuidad con lo que estaba pasando antes. Pasa una cosa que no tiene ni pies ni cabeza, ja, ja, ja, y ¿cómo le llamamos a eso? Pues ¡singularidad! Ale.

Me parece que los visionarios no se imaginan esa singularidad tecnológica como una inteligencia que es cada vez más inteligente, sino como un punto en el que da un salto. Se sale del control. Explota. Y toma el control. No sé si eso es bueno o malo para entender cómo funciona la IA real, la que tenemos ahora. Un buen tema de conversación :-)

Jabuti me dijo que esto de la IA es como la ropa *prêt-à-porter*. El vestido de la modista, hecho a medida, cae mucho mejor, pero es muy caro. Al final todas llevamos ropa hecha en serie, de malas maneras. Y eso va a pasar con la IA. Textos a mano, traducciones a mano, imágenes, música hecha a mano serán mucho mejores, pero no lo podremos pagar y nos rodearemos de cultura hecha por IA. Tú, como soprano, ¿cómo lo ves?

Te voy a pasar más capítulos del libro y me vas diciendo qué te parecen, que me sirve mucho. Escuché a Almudena Hernando en un pódcast de *Sapiens*, de Radio Nacional. Tomé ideas interesantes. Te lo cuento cuando nos demos un paseo.

Muchas ganas de verte.

5. Por fuera

Protégeme

La mayoría de los países europeos utilizan sistemas de evaluación del riesgo como componentes relevantes de sus políticas contra la violencia de género. Los principales métodos de esos sistemas son el juicio profesional no estructurado, el enfoque actuarial y el juicio profesional estructurado.

El juicio profesional no estructurado es un método intuitivo que se fundamenta en la experiencia del profesional, que realiza una evaluación sin utilizar ninguna herramienta estructurada. Se considera un método que introduce mucha subjetividad, por lo que, para superar este inconveniente, surgió el enfoque actuarial, que se basa en ponderar una lista de factores de riesgo a la que se aplican modelos estadísticos y cálculos matemáticos con los que el sistema ofrece una estimación numérica de la violencia futura sin contemplar aspectos cualitativos. Se le critica que no tiene en cuenta factores contextuales. Como mejora del método actuarial surgió el del juicio profesional estructurado, en el que los profesionales utilizan guías que especifican la forma en que se debe obtener y recoger la información que se evaluará para la toma de decisiones sobre el riesgo. Es un método flexible y dinámico, en el que se identifican factores de riesgo comunes, que no se ponderan, a los que el profesional puede añadir otros que considere pertinentes.

En 2007 el Ministerio del Interior del gobierno de España puso en marcha el Sistema de Seguimiento Integral en Casos de Violencia de Género (Sistema VioGén) para evaluar si una mujer víctima de violencia de género sufre riesgo de volver a sufrir esa violencia, es

decir, para predecir la probabilidad de que el agresor reincida con la misma mujer.

Se trata de un algoritmo predictivo de enfoque actuarial basado en métodos estadísticos, pero no en aprendizaje automático, que se aplica cuando la mujer denuncia. Evalúa el riesgo según la suma ponderada de todas las respuestas en función de los pesos preestablecidos para cada variable. En Euskadi y en Catalunya se utilizan otras herramientas de valoración del riesgo propias de esas comunidades autónomas, que también son actuariales.

En el momento de la denuncia los o las agentes de policía formulan a la denunciante las 35 preguntas del Formulario de Valoración Policial (VPR, pp. 177 y 178). Las respuestas, que pueden ser sí o no, aunque algunas se pueden matizar con los valores «leves», «graves» o «muy graves», se introducen en VioGén y el sistema estima automáticamente dos cosas. Por una parte, con un algoritmo calcula el nivel de riesgo de reincidencia del agresor y da un resultado: «no se aprecia», «bajo», «medio», «alto» o «extremo». Por otra, con otro algoritmo, calcula el riesgo de una agresión letal y da un resultado: «bajo» o «alto». En caso de que el riesgo letal sea alto, el algoritmo eleva el valor del riesgo de reincidencia, que es el que se mostrará a los o las agentes; una alerta que indicará que el asunto es de especial interés, para que la protección policial se pueda adecuar a las circunstancias específicas de ese caso, e informará automáticamente al juzgado y a la fiscalía. A partir de esa evaluación de riesgo, la policía establece los mecanismos de protección para la mujer, que están protocolizados por nivel con medidas obligatorias y medidas complementarias.

Preceptivamente, se realiza una reevaluación periódica del riesgo mediante el formulario de Valoración Policial de Evolución del Riesgo (VPER), que tiene dos versiones: VPER-C si ha habido reincidencia y VPER-S si no la ha habido.

Tras un periodo de diez años sin ninguna nueva incidencia, VioGén pasa el caso a inactivo, aunque se puede volver a activar en cualquier momento. Y el caso se da de baja, es decir, se borran los datos, cuando el presunto agresor ha recibido una sentencia firme absolutoria o ha habido un auto de sobreseimiento o de archivo firme, o ha habido una sentencia condenatoria firme, el condenado ha cumplido la pena y ha transcurrido el plazo legal para la cancelación de antecedentes penales.

1. HISTORIA DE VIOLENCIA EN LA RELACIÓN DE PAREJA	Respuestas		
Indicador 1: Violencia psicológica (vejaciones, insultos y humillaciones)	**SÍ**	**NO**	**N/S**
1.1 Intensidad de la violencia psicológica	Leve	Grave	Muy grave
Indicador 2: Violencia física	**SÍ**	**NO**	**N/S**
2.1 Intensidad de la violencia física	Leve	Grave	Muy grave
Indicador 3: Sexo forzado	**SÍ**	**NO**	**N/S**
3.1 Intensidad de la violencia sexual	Leve	Grave	Muy grave
Indicador 4: Empleo de armas u objetos contra la víctima	**SÍ**	**NO**	**N/S**
4.1 Arma blanca 4.2 Arma de fuego 4.3 Otros objetos			
Indicador 5: Existencia de amenazas o planes dirigidos a causar daño a la víctima	**SÍ**	**NO**	**N/S**
5.1 Intensidad de las amenazas	Leve	Grave	Muy grave
5.2 Amenazas de suicidio del agresor	**SÍ**	**NO**	
5.3 Amenazas de muerte del agresor dirigidas a la víctima	**SÍ**	**NO**	
Indicador 6: En los últimos seis meses se registra un aumento de la escalada de agresiones o amenazas	**SÍ**	**NO**	**N/S**
2. CARACTERÍSTICAS DEL AGRESOR			
Indicador 7: En los últimos seis meses, el agresor muestra celos exagerados o sospechas de infidelidad	**SÍ**	**NO**	**N/S**
Indicador 8: En los últimos seis meses, el agresor muestra conductas de control	**SÍ**	**NO**	**N/S**
Indicador 9: En los últimos seis meses, el agresor muestra conductas de acoso	**SÍ**	**NO**	**N/S**
Indicador 10: Existencia de problemas en la vida del agresor en los últimos seis meses	**SÍ**	**NO**	**N/S**
10.1 Problemas laborales o económicos	**SÍ**	**NO**	
10.2 Problemas con el sistema de justicia	**SÍ**	**NO**	
Indicador 11: En el último año el agresor produce daños materiales	**SÍ**	**NO**	**N/S**
Indicador 12: En el último año se registran faltas de respeto a la autoridad o a sus agentes	**SÍ**	**NO**	**N/S**
Indicador 13: En el último año agrede físicamente a terceras personas y/o animales	**SÍ**	**NO**	**N/S**
Indicador 14: En el último año existen amenazas o desprecios a terceras personas	**SÍ**	**NO**	**N/S**
Indicador 15: Existen antecedentes penales y/o policiales del agresor			
Indicador 16: Existen quebrantamientos previos o actuales (cautelares o penales)			
Indicador 17: Existen antecedentes de agresiones físicas y/o sexuales	**SÍ**	**NO**	**N/S**
Indicador 18: Existen antecedentes de violencia de género sobre otra/s pareja/s			
Indicador 19: Presenta problemas un trastorno mental y/o psiquiátrico	**SÍ**	**NO**	**N/S**
Indicador 20: Presenta ideas o intentos de suicidio	**SÍ**	**NO**	**N/S**
Indicador 21: Presenta algún tipo de adicción o conductas de abuso de tóxicos (alcohol, drogas y fármacos)	**SÍ**	**NO**	**N/S**
Indicador 22: Presenta antecedentes familiares de violencia de género o doméstica	**SÍ**	**NO**	**N/S**
Indicador 23: El agresor tiene menos de 24 años	**SÍ**	**NO**	**N/S**

3. FACTORES DE RIESGO/VULNERABILIDAD DE LA VÍCTIMA	Respuestas		
Indicador 24: Existencia de algún tipo de discapacidad, enfermedad física o psíquica grave	**SÍ**	**NO**	**N/S**
Indicador 25: Victima con ideas o intentos de suicidio	**SÍ**	**NO**	**N/S**
Indicador 26: Presenta algún tipo de adicción o conductas de abuso de tóxicos (alcohol, drogas y fármacos)	**SÍ**	**NO**	**N/S**
Indicador 27: Carece de apoyo familiar o social favorable	**SÍ**	**NO**	**N/S**
Indicador 28: Víctima extranjera	**SÍ**	**NO**	
4. CIRCUNSTANCIAS RELACIONADAS CON LOS MENORES			
Indicador 29: La víctima tiene a su cargo menores de edad	**SÍ**	**NO**	**N/S**
Indicador 30: Existencia de amenazas a la integridad física de los menores	**SÍ**	**NO**	**N/S**
Indicador 31: La víctima teme por la integridad de los menores	**SÍ**	**NO**	**N/S**
5. CIRCUNSTANCIAS AGRAVANTES			
Indicador 32: La víctima ha denunciado a otros agresores en el pasado	**SÍ**	**NO**	**N/S**
Indicador 33: Se han registrado episodios de violencia lateral recíproca	**SÍ**	**NO**	**N/S**
Indicador 34: La víctima ha expresado al agresor su intención de romper la relación hace menos de seis meses	**SÍ**	**NO**	**N/S**
Indicador 35: La víctima piensa que el agresor es capaz de agredirla con mucha violencia o incluso matarla	**SÍ**	**NO**	**N/S**

Cuestionario VPR versión 5.0-H. Medidas obligatorias y complementarias según el nivel de riesgo.

Al ser un sistema actuarial automatizado, VioGén, que calcula estadísticamente factores de riesgo predeterminados, hace evaluaciones de riesgo estandarizadas y objetivables, independientemente de la capacitación o pericia del evaluador humano. Su método es mecánico y algorítmico, basado en supuestos lineales, y no permite captar información específica del contexto. No utiliza aprendizaje automático.

Respecto a si VioGén es o no es una inteligencia artificial, teniendo en cuenta la definición popular, la pregunta sería cuánta inteligencia humana es necesaria para predecir si un agresor va a reincidir. Desde mi punto de vista, mucha. Así que un sistema algorítmico que emule esa inteligencia, independientemente del algoritmo con el que esté programado, a mi entender y a efectos de situar un pensamiento crítico, es una inteligencia artificial. Aunque, claro, cabría argumentar que, si la emula mal, entonces no es inteligencia.

Nivel de riesgo	Medidas obligatorias	Medidas complementarias
NO APRECIADO	▪ Las mismas medidas, de tipo operativo y asistencial, que para cualquier otro ciudadano denunciante. Especialmente, información de derechos y de recursos que tiene a su disposición. ▪ Facilitar recomendaciones en medidas de autoprotección	▪ Facilitar a la víctima teléfonos de emergencia y asistencia especializada.
BAJO	▪ Facilitar a la víctima números de teléfono de contacto permanente (24 horas) con las Fuerzas y Cuerpos de Seguridad más próximas. ▪ Contactos telefónicos esporádicos con la víctima. ▪ Comunicación al agresor de que la víctima dispone de un servicio policial de protección. Recomendaciones sobre autoprotección y modos de evitar incidentes. ▪ Información precisa sobre el servicio de teleasistencia móvil. ▪ Derivación de la víctima hacia los servicios sociales y asistenciales que correspondan a su domicilio, recomendándole encarecidamente que se informe de los recursos a su disposición, especialmente los que tengan que ver con su seguridad: puntos de encuentro, viviendas de acogida, etc. ▪ Informar a la víctima sobre las recomendaciones que, para este nivel de riesgo, se establecen en el diseño del Plan de Seguridad del adjunto II. ▪ Si el agresor tiene licencia de armas, requerirle para que las entregue voluntariamente al cuerpo policial actuante. Posteriormente requerir orden a la autoridad judicial competente para la retirada del permiso de armas.	▪ Contactos personales, esporádicos y discretos, con la víctima (acordar con ella la conveniencia de emplear o no uniforme y/o vehículos con distintivos). ▪ Confección de una ficha con los datos relevantes de la víctima y del agresor, que llevará el personal de patrulla. ▪ Acompañamiento al denunciado a recoger enseres en el domicilio, si la autoridad judicial acuerda su salida del mismo.
MEDIO	▪ Vigilancia ocasional y aleatoria en domicilio y lugar de trabajo de la víctima, así como en entrada/salida centros escolares de los hijos. ▪ Acompañamiento a la víctima en actuaciones de carácter judicial, asistencial o administrativo, cuando se considere que puede existir algún tipo de riesgo para la propia víctima. ▪ Entrevista personal con la víctima por el responsable o por personal de la unidad policial encargada de su protección. ▪ Informar a la víctima sobre las recomendaciones que, para este nivel de riesgo, se establecen en el diseño del Plan de Seguridad del adjunto II.	▪ Comprobación periódica del cumplimiento por el agresor de las medidas judiciales de protección. ▪ Entrevista con personal de Servicios Asistenciales que atienden a la víctima/Puntos de Atención Municipal, para identificar otros modos efectivos de protección. ▪ Traslado de la víctima para ingreso en un centro de acogida.

ALTO	▪ Vigilancia frecuente y aleatoria en domicilio y lugar de trabajo de la víctima, así como en entrada/salida centros escolares de los hijos. ▪ Si no lo ha hecho, insistir a la víctima en su traslado a un centro de acogida o al domicilio de un familiar durante los primeros días, especialmente si no se ha procedido a la detención del autor. ▪ Instar el seguimiento obligatorio del agresor mediante dispositivos electrónicos. ▪ Control esporádico de los movimientos del agresor. ▪ Informar a la víctima sobre las recomendaciones que, para este nivel de riesgo, se establecen en el diseño del Plan de Seguridad del adjunto II.	▪ Contactos esporádicos con personas del entorno del agresor y de la víctima: vecinos, familia, trabajo, lugares de ocio...
EXTREMO	▪ Vigilancia permanente de la víctima, hasta que las circunstancias del agresor dejen de ser una amenaza inminente. ▪ Control intensivo de los movimientos del agresor, hasta que deje de ser una amenaza inminente para la víctima. ▪ En su caso, vigilancia en entrada/salida centros escolares de los hijos. ▪ Diseño de un plan de seguridad personalizado para cada víctima, sobre las medidas que, para este nivel de riesgo, se establecen en el catálogo del Plan de Seguridad del adjunto II.	

Formulario VPER-C 4.0 para la evaluación del riesgo cuando ha habido reincidencia.

Afirmar que un modelo estadístico es inteligencia artificial es controvertido. Para las lógicas de mercado actuales no lo es, pero para las conceptualizaciones académicas sí lo es. Como vimos, el reciente reglamento de la Unión Europea evita definir la inteligencia artificial por las técnicas con las que se programa, así que habrá que ver cómo se va decantando su aplicación.

Si no se tuviera información sobre el modelo estadístico de VioGén, información que por otra parte no es tanta, puesto que su código no es público, viendo solo su funcionamiento, por fuera, ¿no podría considerarse una inteligencia artificial, quizás de gama baja, quizás algo tosca? De hecho, en varias ocasiones el Ministerio del Interior ha anunciado su apuesta por incorporar analíticas más avanzadas, algoritmos de aprendizaje automático, y no sería de extrañar que se incluyan en un futuro cercano.

Formulario VPER-Valoración Policial de Evolución del Riesgo (CON INCIDENTES)			
Fuentes de información Víctima ☑ Agresor ☑ Testigo(s) ☑ Otras (informes técnicos, médicos, etc.) □			
P01-¿Ha existido algún tipo de violencia por parte del agresor desde la última valoración?	Sí ☑	No	No se sabe
101. Vejaciones, insultos, humillaciones	Sí ☑	No	No se sabe
	Leves ☑	Graves	Muy graves
102. Violencia física	Sí ☑	No	No se sabe
	Leves ☑	Graves	Muy graves
103. Violencia sexual	Sí ☑	No	No se sabe
	Leves ☑	Graves	Muy graves
P02.-¿Ha empleado el agresor armas u objetos contra la víctima desde la última valoración?	Sí ☑	No	No se sabe
105. El agresor empleó Arma blanca ☑ Arma de fuego □ Otros objetos □			
106. ¿Tiene acceso a armas de fuego a través de terceros?	Sí ☑	No	No se sabe
P03-¿La víctima recibe o ha recibido amenazas o planes dirigidos a causar daño físico/psicológico desde la última valoración?	Sí ☑	No	No se sabe
	Leves ☑	Graves	Muy graves
De suicidio por parte del agresor □ Económico-materiales ☑ De muerte □ A la reputación social □ A la integridad y/o custodia de los hijos □			
P04. Incumplimiento de disposiciones judiciales cautelares o quebrantamiento de penas o medidas penales de seguridad desde la última valoración	Sí ☑	No	
El agresor se ha puesto en contacto por vía telemática con la víctima	Sí ☑	No	
El agresor se ha puesto en contacto con la víctima a través de terceros	Sí ☑	No	
El agresor se ha acercado a la víctima	Sí ☑	No	
P05. Celos exagerados, control y/o acoso desde la última valoración	Sí ☑	No	No se sabe
109. El agresor muestra celos exagerados sobre la víctima o tiene sospechas de infidelidad	Sí ☑	No	No se sabe
110. El agresor muestra conductas de control sobre la víctima	Sí ☑	No	No se sabe
Físico (limitación de movimientos) □ Psicológico y/o social ☑ Escolar / laboral □ Económico Cibernético (controla redes sociales, mensajes, llamadas, contactos) □			
111. El agresor muestra conductas de acoso sobre la víctima	Sí ☑	No	No se sabe
P06. El agresor está fugado o en paradero desconocido	Sí ☑	No	
P07. Evidencias de comportamientos por parte del agresor desde la última valoración	Sí ☑	No	No se sabe
113. Se ha distanciado de la víctima	Sí ☑	No	
114. Muestra una actitud pacífica, asume su situación con respeto a la víctima, sin ánimo de venganza contra ella ni su entorno	Sí ☑	No	
115. Exterioriza una actitud respetuosa hacia la Ley y de colaboración con los agentes	Sí ☑	No	
116. Muestra arrepentimiento	Sí ☑	No	No se sabe
117. Se acoge a programas de ayuda	Sí ☑	No	No se sabe
118. Cumple con el régimen de separación y cargas familiares	Sí ☑	No	No procede

VioGén se enmarca dentro del uso de lógicas actuariales algorítmicas para la toma de decisiones automatizada, del mismo modo que lo son el algoritmo Correctional Offender Management Profiling for Alternative Sanctions (COMPAS), utilizado por los tribunales estadounidenses para decidir cárcel o libertad mientas el acusado está a la espera de juicio, o VeriPol, empleado por la policía española para determinar si una denuncia es falsa. COMPAS se entrenó con aprendizaje automático, por lo que claramente es una inteligencia artificial. En VeriPol se utilizan técnicas de procesamiento del lenguaje natural y aprendizaje automático, por lo que es evidente que también lo es. Más allá del algoritmo concreto, este tipo de herramientas actuariales, por sus sesgos conocidos, está en el punto de mira de las organizaciones defensoras de la justicia social. Por eso, en un contexto de reflexión sobre la inteligencia artificial, en mi opinión tiene sentido pararse a ver cómo el algoritmo VioGén se ensambla en la lucha tecnosocial contra la violencia de género.

En el año 2022 Eticas Foundation, cuya misión es proteger a las personas y al medio ambiente en los procesos de inteligencia artificial sesgados, inició una auditoría externa de VioGén. Con sus auditorías, Eticas pretende generar datos empíricos y perspectivas sobre las implicaciones sociales y éticas de las tecnologías de inteligencia artificial que puedan influir en las decisiones políticas, y abogar por medidas reguladoras que den prioridad a la equidad, la responsabilidad y la justicia social. Esta auditoría se realizó sin acceso al código y sin acceso al conjunto de datos con el que el algoritmo fue construido y validado, ya que el Ministerio del Interior ni los ha liberado ni permite que investigadores independientes puedan analizarlos. La falta de transparencia de VioGén es evidente.

Cuando ni el código ni los datos de validación están disponibles, una auditoría consiste en abrir una caja negra de la cual se conocen sus efectos fuera, pero se desconoce su funcionamiento por dentro. En este caso se aplicó una metodología cualitativa y cuantitativa. Para el análisis cuantitativo se partió del registro público de casos documentados por el Consejo General del Poder Judicial, única fuente de datos pública disponible. Para el análisis cualitativo se realizaron entrevistas telefónicas semiestructuradas a 31 mujeres que habían pasado por VioGén. También se entrevistó a representantes de la Fundación Ana Bella, Red Global de Mujeres Supervivientes, para conocer la perspectiva de la sociedad civil. Y se pasaron cuestionarios *online* a 7 abogados o abogadas especialistas en violencia de género.

El objetivo principal de la auditoría era escrutar la existencia de sesgos en la prevalencia del homicidio entre los diferentes grupos protegidos. En este caso no se buscaba una discriminación sociocultural o étnica, sino que se quería estudiar la presencia de sesgos entre los grupos que genera el algoritmo: riesgo que no se aprecia, bajo, medio, alto o extremo.

Se observaron los falsos negativos, es decir, casos en los que las mujeres asesinadas habían denunciado, pero no recibieron protección porque el sistema calificó su riesgo como «no apreciado» o «bajo». El 56 % de las asesinadas cayeron en esta categoría. Se observaron casos de protección insuficiente, es decir, aquellos en los que las mujeres asesinadas estaban recibiendo protección, pero esta no consiguió evitar el asesinato. Estos casos son un subconjunto de los verdaderos positivos, los que el sistema identificó con riesgo de reincidencia y efectivamente la hubo. El 44 % de las asesinadas cayeron en esa categoría.

Todo modelo de predicción aplicado a lo social plantea la cuestión no técnica, sino política, de cuántos falsos positivos (hombres que se predijo que serían reincidentes y por tanto pudieron ver limitadas sus libertades sin serlo) y cuántos falsos negativos (hombres que se predijo que no serían reincidentes y lo fueron) la sociedad está dispuesta a asumir. Esto teniendo en cuenta que un modelo estadístico da un valor predictivo, calcula una probabilidad, no una certeza. Y, como oí decir a una psiquiatra, en lo referente al malestar del otro, los modelos son herramientas, nunca verdades.

En principio VioGén es un sistema de recomendación, es decir, un auxiliar en la toma de decisiones. Los o las agentes de policía, a su criterio, pueden aumentar la puntuación de riesgo (aunque en ningún caso disminuirla). Sin embargo, según Eticas, la policía solo modifica al alza esa puntuación en un 5 % de los casos. Para el otro 95 % aplica automáticamente la valoración del algoritmo. Esto ocurre porque se produce un sesgo de autoridad tecnológica, también llamado de automatización.

Aunque una sentencia de la Audiencia Nacional recuerda que «la respuesta policial a la violencia contra la mujer exige que el sistema pueda prevenir la violencia y reevaluar el riesgo, esto es, más allá de la recogida de datos automatizados, la predicción y la prevención son la finalidad primordial del sistema de evaluación que exige agentes especializados en su tratamiento y sensibilización en su seguimiento», el

hecho es que disponer de un algoritmo genera una confianza muy alta en sus predicciones, confianza debida o bien a ese sesgo de autoridad tecnológica o bien al hecho de que las personas agentes de policía no se sientan lo suficientemente expertas, no se consideren una autoridad en la materia con la seguridad y el aplomo suficientes como para corregir el algoritmo. Así que, aunque concebido como una herramienta auxiliar, en la práctica las decisiones se delegan en él.

Respecto a la recogida de datos, la declaración de la denunciante no se introduce íntegra en el sistema. El o la agente de policía, para cada pregunta-indicador, tiene que traducir la declaración de la mujer en una fórmula mucho más simple y sin matizaciones: «sí» o «no». Además, la denunciante suele ser la única fuente de información sobre el agresor. Es ella quien señala las características de este. Dependiendo de sus circunstancias socioculturales, puede no estar entendiendo bien las preguntas o la relevancia, las consecuencias de cada respuesta. O puede estar en un estado emocional alterado y a causa de ello no organizar bien sus ideas o no recordar bien todos los detalles de los hechos pasados. O puede tener una autoestima baja y no caracterizar adecuadamente el riesgo que el agresor presenta sobre ella misma. La denunciante puede tener creencias o patologías que tiendan a justificar la violencia que está sufriendo. Es posible que durante la entrevista no esté acompañada por ninguna persona de su confianza ni tenga ayuda legal. Puede tener miedo, vergüenza, vacilar… Es cierto que, además del peritaje policial, existe la posibilidad de hacer una valoración forense del agresor en sede judicial, pero en la práctica esto apenas ocurre y las medidas de protección de la víctima se adoptan teniendo en cuenta únicamente el informe policial.

Por otra parte, no está claro que haya un perfil psicológico del hombre agresor, es decir, que los agresores constituyan un grupo homogéneo. Otras investigaciones señalan que no todos los indicadores del cuestionario útiles para predecir la reincidencia lo son para predecir futuros asesinatos. La violencia letal y la no letal pueden ser fenómenos diferentes, aunque ambas se den en el marco de la violencia de género. El perfil del potencial feminicida no es el perfil medio del maltratador.

Respecto al cuestionario, no está clara la relevancia de algunos indicadores, como por ejemplo si tiene menores a su cargo (sean hijos o no). El hecho de no tenerlos influye de manera negativa en la evaluación del riesgo. Según Eticas, las mujeres asesinadas que no tenían menores a su cargo habían recibido por sistema una

puntuación de riesgo más baja que las que sí los tenían, con una diferencia en torno al 44 %.

La auditoría indica que VioGén adapta las evaluaciones de riesgo a los recursos disponibles. Esto significa que el sistema califica con riesgo extremo solo el número de casos que puede permitirse, por lo que los recortes de financiación tienen un impacto directo cuantificable en las posibilidades de que las mujeres reciban protección policial. Es decir, la puntuación del algoritmo no solo viene determinada por las respuestas introducidas en el cuestionario, sino también por los recursos disponibles, que dependen de la capacidad policial, pero también de la distribución global de las mujeres denunciantes a proteger.

Llevado a cifras, en 2021 solo una de cada siete mujeres que pidieron protección policial la recibió. Se calcula que solo el 21 % de mujeres mayores de 16 años víctimas de violencia de género ponen denuncia, es decir, entran en VioGén, lo cual significa que solo el 3 % de la totalidad de mujeres que sufren violencia de género fueron calificadas con un riesgo de nivel medio o superior y, en consecuencia, recibieron protección policial.

En el año 2020 fueron asesinadas cincuenta mujeres. Solo ocho habían puesto denuncia. En el 2021, cuarenta y nueve. Diez habían puesto denuncia. En el 2022, cuarenta y nueve. Veinte habían denunciado. En el 2023, cincuenta y cinco. Catorce habían denunciado.

Siempre según Eticas, en el 45 % de los casos VioGén clasifica el riesgo como «no se aprecia». En las entrevistas las mujeres manifestaron que el hecho de ir a la policía a poner una denuncia contra el agresor ya es una acción de riesgo que puede acarrear represalias. Cuando el caso recibe una calificación de «no se aprecia» o «bajo», se crea un brecha muy grande entre cómo la mujer autopercibe su situación de riesgo y cómo la percibe el algoritmo. Las mujeres se sienten engañadas.

Sin embargo, la política estatal contra la violencia de género hace de la denuncia el eslabón fuerte de la cadena. Si no hay denuncia, no hay visibilidad, no hay caso. No hay posibilidad de protección. El gobierno de España califica la violencia de género como un problema estructural, un problema de Estado cuya política de solución pasa principalmente por la denuncia, bien sea de la propia mujer, bien sea de su entorno, lo cual permite activar el «protocolo cero», por el cual la policía investiga el caso sin necesidad de denuncia de la víctima.

Pero, sea como sea, debe haber denuncia. «La denuncia es la que nos permite abrir el paraguas protector del Estado español». Está claro que el éxito de esta política de denuncias requiere que las víctimas y la sociedad en su conjunto superen enormes barreras emocionales, estructurales e institucionales y tengan una confianza máxima en las Fuerzas y Cuerpos de Seguridad del Estado y en el sistema algorítmico de evaluación del riesgo.

En resumen, en la política que el gobierno de España ha establecido para luchar contra la violencia de género, considerada un problema de Estado, la denuncia es la clave de bóveda.

Aproximadamente, solo una de cada cinco mujeres que sufren violencia de género pone denuncia.

Aunque VioGén nominalmente es un asistente para la evaluación policial del riesgo, en la práctica la decisión se delega en el algoritmo.

El algoritmo no se puede auditar, porque el Ministerio del Interior no permite el acceso al código ni a los datos con los que se ha construido y validado el modelo.

El algoritmo carece de transparencia.

Establecer si VioGén es o no es una inteligencia artificial es de gran trascendencia, puesto que, si lo es, debería cumplir los criterios de transparencia, explicabilidad y trazabilidad que establece el reciente Reglamento Europeo para la Inteligencia Artificial.

La valoración que hace el algoritmo no depende solo de los datos del caso, sino también de los recursos disponibles.

El impacto de algunos indicadores es discutible.

La recogida de datos puede ser muy parcial, incompleta o sesgada. No hay garantías de que no lo sea.

En la construcción del algoritmo no se ha tenido en cuenta a las destinatarias. No se ha consultado a mujeres víctimas de violencia de género ni a sus asociaciones.

Y a todo esto, más allá de la auditoría de Eticas, está la percepción generalizada de que las políticas contra la violencia de género no están funcionando. Ahora bien, ¿qué parte de este no funcionamiento se puede atribuir al modelo estadístico? ¿Y al procedimiento de recogida de datos? ¿Y a los presupuestos económicos y a los recursos policiales disponibles? ¿Y a la formación de las o los policías

o guardias civiles que procesan la denuncia? ¿Y qué parte se puede atribuir al machismo estructural? ¿Y a la confianza ciudadana en las Fuerzas y Cuerpos de Seguridad del Estado?

Si miramos solo al algoritmo, el camino es mejorarlo. En efecto, se habla de transformarlo introduciendo técnicas de aprendizaje automático, aunque ello plantea problemas. El primero es que, por la naturaleza del asunto, ni legal ni éticamente se podría establecer un grupo de control para medir el impacto del modelo. El segundo es que los datos históricos sobre la reincidencia van a reflejar la reincidencia ocurrida después de que se tomaran medidas policiales. Pero esa reincidencia ocurrida no refleja el riesgo de reincidencia cuando no se toman medidas. El modelo podrá aprender a partir de los datos de reincidencia ocurrida, pero no sobre el riesgo real de reincidencia, porque el algoritmo es recursivo: con sus decisiones incide y transforma la realidad que debe modelizar. Aun superando esos problemas, cabe reflexionar sobre cuál sería el impacto de introducir el aprendizaje automático en el algoritmo sobre la lucha global contra la violencia de género si no hay ninguna otra modificación ni en los indicadores, ni en la recogida de datos, ni en los recursos disponibles…

Atendiendo al algoritmo, el campo de batalla que se abre es el de la responsabilidad algorítmica: cómo y quién responde por los daños ocasionados, qué mecanismos de revisión e impugnación se ofrecen, cómo se ha testado y validado el modelo… Atendiendo al todo en el que el algoritmo es solo una pieza más, un posicionamiento crítico orientaría la lucha hacia las fuerzas estructurales, las relaciones de poder, la gobernanza del conjunto del sistema…

Cierto que todo algoritmo se puede mejorar, así como también se puede descartar. Pero ¿qué pasa con todo lo demás? Ningún asunto algorítmico es solo algorítmico.

En el momento de terminar estas líneas, se está poniendo en marcha VioGén 2, que introduce cambios en el sistema. A propósito de estos cambios, Paz Lloria, catedrática de Derecho Penal de la Universitat de València y experta en violencia de género, en una entrevista radiofónica en el programa de radio *Las mañanas de RNE* dice:

> Sería importante entrenar al algoritmo en perspectiva de género. Muchas veces nos olvidamos de que las decisiones automatizadas están introducidas por personas humanas. Hay que introducir la perspectiva de género en la valoración de los datos, porque la valoración del algoritmo al final es una valoración que ha sido introducida con unos

> criterios humanos que, si no cuentan con esa perspectiva, es difícil que valoren determinadas situaciones. Esto también se puede corregir si después del resultado que ofrece el sistema la propia policía corrige o matiza esa valoración. El problema no es el posible sesgo machista del algoritmo, sino la falta de sesgo de género.

Y explica que poner sesgo de género consiste en valorar ítems que no están contemplados. Está utilizando la palabra «sesgo» como sinónimo de «discriminación positiva».

Cuídame

En Las Vegas, un destino turístico lleno de hoteles, zonas comerciales y casinos, se especula con que en 2035 entre el 38 % y el 65 % de los puestos de trabajo de servicios podrían estar automatizados. Las empresas de turismo y hostelería quieren reducir costes laborales. Si con inteligencia artificial se puede sustituir a trabajadores sin que ello afecte a la productividad, los beneficios o la experiencia del cliente, se hará. Y se está haciendo. Cajeros automáticos sustituyen a recepcionistas, robots de texto recomiendan restaurantes, automatismos sirven cócteles detrás de la barra…

En mayo de 2024 el sindicato Culinary Union, que agrupa a camareras de piso, camareros de hostelería, porteros, botones, cocineros, lavanderas y trabajadores de cocina principalmente latinos, negros, asiáticos y del Pacífico, declaró una huelga en Virgin Las Vegas, la única cadena de hoteles que se negaba a firmar el nuevo convenio. La lucha era, entre otras cosas, por pasar de 26 a 35 dólares la hora.

Este sindicato también lucha por hacer frente a la desaparición de puestos de trabajo. En anteriores convenios ha conseguido que, cuando un puesto de trabajo se va a eliminar por automatización, se le notifique con antelación a la trabajadora; que se aumente la remuneración de méritos y se incrementen las cotizaciones a la atención médica y a los fondos de pensiones para las trabajadoras que son despedidas a raíz de la introducción de nueva tecnología; que se realicen formaciones para los nuevos puestos de trabajo creados a partir de esta tecnología; que se reconozca el derecho a negociar sobre la tecnología que rastrea la ubicación de las empleadas; que se notifique

y se reconozca la posibilidad de negociar con respecto al intercambio de datos personales; y que se establezca el derecho a la indemnización para las empleadas que reciben propinas por sus servicios si la infraestructura tecnológica necesaria falla e impide a la empleada realizar su trabajo.

En Londres, el colegio privado David Game College ha puesto en marcha una prueba piloto en la que el alumnado de entre 15 y 17 años aprende a través de pantallas y gafas de realidad virtual, sin profesorado físico en el aula. Cada estudiante sigue un plan de estudios personalizado. La inteligencia artificial monitorea sus progresos y le ofrece nuevos contenidos y actividades en materias tradicionales como matemáticas, lengua, idiomas, geografía, historia, ciencias, etcétera. El alumnado está en un aula clásica, en la que hay unas personas adultas que vigilan y prestan apoyo, pero no enseñan. Es una escuela sin profesorado, excepto en educación física, arte, valores o sexualidad, que sí son impartidas por humanos. El programa tiene un coste anual de unos 32.000 euros y quiere solucionar el déficit de profesionales, pues por lo visto cuesta encontrar personas que quieran ejercer como docentes. Y también pretende mejorar la experiencia del estudiantado en las aulas.

En la ciudad de Barcelona viven unas 350.000 personas con más de 65 años. De ellas, 90.000 viven solas. Se estima que el 90 % de estas personas quieren quedarse en su casa el máximo tiempo posible y no ir a una residencia o estar en casa de familiares. En enero de 2021, la Tinència d'Alcaldia de Drets Socials, Justícia Global, Feminismes i LGTBI del Ajuntament de Barcelona estableció una medida de gobierno para un nuevo modelo de ciudad: cuidarnos en comunidad. En sus propias palabras: «La pandemia ha puesto de relieve y ha agudizado la crisis de cuidados que ya padecía nuestra sociedad. Se ha visto más claramente que nunca que las mujeres asumen la mayoría de los trabajos de cuidados tanto en el hogar como en instituciones y residencias para personas mayores. Ante esta situación, con esta medida de gobierno nos proponemos activar la innovación social de la mano de la ciudadanía y otros actores económicos y el mundo educativo, cultural y social». El objetivo principal de la medida de gobierno era mejorar las respuestas a los problemas sociales incrementando la calidad, la eficacia y la eficiencia de las políticas sociales municipales.

Una de las actuaciones consistió en aplicar las tecnologías al acompañamiento y los cuidados para mejorar la calidad de vida y los procesos de envejecimiento de las personas mayores, facilitar la

autonomía en el propio domicilio en la primera fase de deterioro cognitivo o dependencia, asegurar la sostenibilidad en la prestación de los servicios públicos y posibilitar un sector industrial 4.0 competitivo en el ámbito internacional. La actuación ampliaba una prueba piloto iniciada en el año 2020 que ponía en el domicilio de personas mayores dependientes un robot asistente inteligente. La dotación económica era de 180.000 € para 50 robots. Estos se agregaban a otros 50 ya en funcionamiento en una prueba piloto anterior, de modo que se llegaría a una cifra de 100, con el objetivo de obtener una muestra relevante y conseguir más pruebas y más conocimiento antes de desplegar el proyecto. El piloto inicial se había desarrollado en el marco de la convocatoria del reto «Cómo mejorar la calidad de vida de las personas mayores mediante la tecnología», lanzado por la fundación Mobile World Capital Barcelona. Los robots, SOMCARE, estaban fabricados por la empresa leridana Grupo Saltó, que trabaja tanto para el sector público como para el privado.

Desde el Ajuntament de Barcelona, el proyecto se presentaba como una teleasistencia avanzada y móvil. El robot convive con la persona mayor, se comunica con ella por voz, se desplaza autónomamente dentro de la vivienda e incorpora inteligencia artificial para adaptarse a sus necesidades específicas. Un complemento a los cuidados familiares y profesionales.

Para el Grupo Saltó, la mayor innovación consistía en conectar los sensores del robot con su plataforma SOM, usando Internet de las cosas (IoT) y comunicaciones 5G. El robot, en conexión con la plataforma, se puede configurar para hacer llamadas telefónicas a familiares, generar alertas, tomar fotografías o vídeos, recordar la toma de medicación o las visitas médicas, tomar la temperatura, sugerir pautas de alimentación, proponer ejercicios de memoria, conversar y, en definitiva, paliar la soledad y cuidar. Familiares y profesionales se pueden conectar con la plataforma para obtener información sobre el estado de la persona mayor y sus incidencias. La inteligencia artificial puede analizar, predecir y ayudar en la toma rápida de decisiones. Se presenta como una humanización de las tecnologías al servicio de los cuidados.

También participaron en el proyecto el Instituto de Robótica para la Dependencia —fundación catalana cuya misión es «mejorar la calidad de vida de las personas dependientes, la de sus familias y la de los profesionales que les ofrecen apoyo por medio de la tecnología más avanzada, impulsando la obtención de productos innovadores que

transformen el estado actual de las cosas, transfiriendo conocimientos, actuando con eficiencia, calidad, seguridad y sostenibilidad»— y la Fundación i2CAT —centro catalán de investigación e innovación que quiere «construir la sociedad del futuro investigando en proyectos estratégicos, fortaleciendo políticas digitales de la administración pública y transfiriendo tecnologías para que las empresas desarrollen soluciones innovadoras orientadas al mercado».

Un par de meses antes, trabajadoras del Servei d'Atenció Domiciliària de Barcelona (SAD) organizaban protestas contra la externalización del servicio. El ayuntamiento había adjudicado a la empresa Servisar el 50 % de los servicios del SAD. Las trabajadoras denunciaban que esta empresa es una multinacional, propiedad de un fondo de inversión, que ha hecho una gestión pésima en las residencias geriátricas. La adjudicación suponía una precarización de las trabajadoras. Reclamaban pararla y remunicipalizar el servicio.

En 2023, con financiación del programa Next Generation EU, la Generalitat de Catalunya anunció que se disponía a adquirir mil robots sociales para asistir y acompañar a personas mayores que vivieran solas. Doscientos de esos robots se los cedería al Estado para que la iniciativa se aplicase también en el resto de España.

El anuncio se producía en un contexto de precarización del sector de las profesionales de la asistencia domiciliaria y los cuidados. Desde el Sindicato de Cuidadoras, vía Twitter, denunciaron: «No fueron capaces de facilitar grúas para levantar a personas en sus domicilios, no fueron capaces de escuchar mensajes sobre la precarización de los cuidados. Sí de gastar el dinero del pueblo en robots».

La asociación Els Estels Silenciats, formada en el contexto de la pandemia de COVID-19 para luchar por los derechos y la dignidad de las personas mayores que viven en residencias, en verano de 2024 denunció que la patronal de los geriátricos catalanes Associació Catalana de Recursos Assistencials (ACRA) viajaría a Japón para valorar la utilización de asistentes tecnológicos ante el aumento demográfico de personas mayores y la falta de personal para atenderlas. La presidenta de ACRA, Cinta Pascual, explicó que no encuentran profesionales que quieran trabajar en las residencias. Una causa es económica: las condiciones laborales. Por ello pedía a la consellera de la Generalitat, Mónica Martínez Bravo, que aceptase las demandas del sector respecto a las mejoras salariales. Otra es vocacional: la sociedad ha perdido la vocación de cuidar. Habría que

situar los cuidados a las personas dependientes como una profesión con voluntad de servicio, estable y con futuro. Sabemos que las personas quieren envejecer en su casa y hay que hacer todo lo posible para facilitarlo. Pero las residencias son necesarias cuando la dependencia se agrava. Y el envejecimiento de la población es inminente.

Los robots sociales para el cuidado de mayores están proliferando. A las personas mayores les cuesta utilizar los asistentes integrados en el teléfono móvil, y los altavoces inteligentes, tipo Siri o Alexa, carecen de retroalimentación visual. Los robots sociales son máquinas físicas que entran en escena como propuestas más eficaces y amigables.

La robótica social es uno de esos campos en los que hay perspectiva de negocio. Se percibe la asistencia personal como un conjunto de tareas en las que hay muchas repeticiones y, por tanto, se puede automatizar. La expectativa es aumentar eficiencia y reducir costes.

Sin embargo, la industria de la robótica social reconoce la existencia de lo que ellos mismos denominan como desafíos: dificultad para la integración efectiva de robots en entornos complejos, necesidad de superación de barreras culturales y normativas, miedo a la falta de ética y pérdida de la privacidad de los datos personales, falta de aceptación y confianza de las personas en la interacción con los robots…

Según un informe de agosto de 2024 de Business Research Insights, organización que ofrece datos sobre mercados, el tamaño del mercado mundial de robots sociales fue de más de quinientos millones de dólares y se prevé que para el 2028 llegue a los mil millones. Los robots sociales se utilizan en cuidados domésticos, asistencia sanitaria y apoyo educativo. La pandemia de COVID-19 tuvo un efecto dual. Por una parte, se ralentizó la producción y la incertidumbre limitó el crecimiento del mercado. Por otra, aumentó la percepción favorable respecto a su utilización, especialmente para los robots humanoides que comprenden emociones, entablan conversaciones, realizan tareas, interaccionan y son empáticos. Pese a ello, hay miedo.

Un robot social es una convergencia entre hardware (sensores, actuadores y procesadores), software (sistemas operativos y algoritmos de control, de reconocimiento visual, de procesamiento del lenguaje natural, de reconocimiento de emociones o de navegación autónoma) y servicios (plataformas, integración, personalización, mantenimiento o reparación). Hacen un uso intensivo de Internet de las cosas (IoT), de las comunicaciones 5G y de la inteligencia artificial.

Por el momento, para comprar o implementar robots se requiere mucha inversión, pero la está habiendo. Los mercados tienen expectativas en los sectores de atención médica, educación y comercio minorista. China aplica un modelo estatal de control social; Estados Unidos, sin trabas regulatorias, un modelo mercantil orientado al beneficio; y Europa parece que se especializaría en la ética.

Según *The Guardian*, en Japón no hay suficientes cuidadores humanos como para satisfacer las necesidades de cuidados de toda la población que los necesita. En 2025 el déficit podría ser de 370.000 cuidadoras, así que el gobierno quiere que aumente la aceptación de los robots para aliviar la carga de trabajo de la fuerza laboral. Los robots pueden ayudar a levantarse de la cama y sentarse en la silla de ruedas, a ir al baño o a caminar por la calle. No tienen por qué ser robots sociales. Pueden ser parecidos a los robots industriales. Las barreras son el coste y el rechazo social. La expectativa del gobierno japonés es que cuatro de cada cinco personas que reciben atención acepten recibir algún tipo de apoyo por parte de robots.

En el Mobile World Congress 2024 se presentó el proyecto ATENEA: inteligencia artificial para el bienestar de las personas mayores. Es un proyecto de colaboración público-privada liderado por una empresa tecnológica (Momentum Analytics) y una entidad social (Grupo ABD) en cuyo diseño han participado personas mayores y que ha recibido un dictamen ético favorable por parte de la comisión de bioética de la Universitat de Barcelona. Según los promotores, ATENEA es un instrumento. Nunca sustituirá a una persona, pero contribuye a mejorar el bienestar de todas estas personas que están excluidas de la sociedad. En el momento de la presentación el proyecto estaba subvencionado por el Departament de Drets Socials de la Generalitat de Cataluña en el marco del Plan de Recuperación, Transformación y Resiliencia, financiado por la Unión Europea, Next Generation EU con un presupuesto de 2.300.000 €.

ATENEA es un asistente de voz para hacer gestiones. Autentifica a las usuarias por biometría de voz. Se puede usar para pedir cita médica, preguntarle cómo llegar a un sitio, conectar por voz o por viodeconferencia con una familiar y cosas así. También se puede usar con un reloj que mide las pulsaciones, detecta caídas y, si nota algo raro, llama a emergencias. Funciona en catalán o castellano y se orienta a las necesidades de las personas mayores o dependientes.

A principios de 2024 se presentó el proyecto Celia. Producido por atlanTTic, Centro de Investigación en Tecnologías de Telecomunicación de la Universidad de Vigo, y financiado por la Xunta de Galicia, es un asistente programado con inteligencia artificial para paliar la soledad no deseada. Es un chatbot que se tiene que usar integrado en Whatsapp. Está disponible para su descarga gratuita en Google Play.

Celia hace compañía: por propia iniciativa, da conversación, informa, propone entretenimientos y va realizando tests neuropsicológicos para estimular y comprobar las habilidades cognitivas de las personas. El objetivo es ayudar a detectar enfermedades como el Alzheimer y la demencia en sus fases tempranas. Evalúa la salud cognitiva a través del análisis de lenguaje y de la voz, asumiendo que de ella es posible extraer indicadores sobre dolencias neurológicas, emocionales o cardiorrespiratorias.

La Universidad Carlos III de Madrid ha desarrollado el robot social experimental Mini. Es un compañero de vida para personas con deterioro cognitivo. Tiene un cuerpo blando con sensores de tacto, para que sea agradable tocarlo y se pueda jugar físicamente con él. Tiene articulaciones y ojos expresivos, así que puede hacer gestos y expresar sentimientos, y también micrófonos y altavoces para interactuar con voz. Es proactivo y puede tomar decisiones. Propone actividades y ejercicios. Puede reproducir noticias de actualidad, configurando fuentes veraces de información, y dar información meteorológica. Detecta el rostro de la persona mayor e interpreta sus expresiones faciales. Detecta la presencia de otras personas. Mediante sensores IoT, monitoriza el comportamiento supervisando sus hábitos de higiene, su calidad del sueño y otros indicadores de depresión. Facilita que la persona mayor se comunique con sus allegadas a través de Whatsapp. Y se registran todas las interacciones entre la persona y el robot: tipo, frecuencia, duración, etcétera.

Mediante la captación de todos esos datos, se realizan las siguientes métricas: la emoción media de la persona, si mantiene relaciones sociales con otras, sus patrones de comportamiento en casa, el tiempo de interacción con el robot y el tiempo de interacción en Whatsapp. Si determinadas frecuencias, configurables, no se cumplen (por ejemplo, interacción física con familiares y amigos al menos una vez al mes), se utilizarán sensores IoT y un chatbot para calcular si hay comportamientos depresivos. Dependiendo del resultado, el robot evalúa el riesgo de depresión. Si el riesgo existe, hace preguntas, propone

actividades de socialización y monitoriza la evolución. Lógicamente, está conectado a Internet.

En este proyecto se reconoce la necesidad de una validación por parte de profesionales de la psicología y los desafíos que plantea el trasiego de datos personales. Cuando el sistema fuera validado, se podría entrenar un modelo de aprendizaje automático para determinar el riesgo de padecer depresión sin necesidad de formular preguntas. Esto permitiría anticipar y prevenir comportamientos relacionados con la soledad con base en los datos recogidos por los detectores.

En Corea del Sur el gobierno ha distribuido siete mil robots con inteligencia artificial para combatir la soledad no deseada de las personas mayores. Es un muñeco de peluche que integra ChatGPT, sensores y telefonía móvil. Un nietecito artificial que monitorea a la abuela y le proporciona ¿amor?

Desde el punto de vista algorítmico, por más que haya una promesa de personalización, ¿cómo van a funcionar robots sociales que interpreten las emociones y los comportamientos de forma lo suficientemente estandarizada con modelos matemáticos como para que su producción sea masiva, cuando sabemos lo singular que es cada una de nuestras personas mayores dependientes?

¿Qué consecuencias puede tener el hecho de que una persona mayor establezca un vínculo emocional con el robot? ¿Qué gobernanza, qué auditorías, qué software libre garantizan que el robot o, más bien, la empresa que está detrás hará un buen uso del poder que tendrá sobre la persona?

En definitiva, sistemas artificiales que no tienen ni moralidad ni conciencia, pero actúan. Humanoides, brazos mecánicos y chatbots coexistiendo con cuidadoras en condiciones de trabajo penosas y precarias. Cóctel de expectativas de mercados y administraciones, aderezados con narrativas de ciencia ficción.

En el siglo pasado, la comunidad psiquiátrica implicada en el movimiento contracultural antipsiquiátrico dejó de considerar la locura como un error. De repente, la locura se humanizó. Tenía algo que decir y había que escucharla. Ahora hay que elaborar qué es lo que la vejez viene a decirnos, para ensamblar automatizaciones que no silencien esa voz.

Ideología

«Inteligencia artificial» es un muy buen término para el *marketing*, pero no tan bueno para pensar o entender.

La definición que circula por la calle es que inteligencia artificial es una tecnología que permite crear máquinas computacionales que emulan la inteligencia humana. Pero esto tiene contornos muy borrosos, pues habría que dibujar un círculo que delimitase qué es lo que entra dentro de la inteligencia humana y qué es lo que no, lo cual es problemático.

Por ejemplo, una calculadora medianamente avanzada, que pueda calcular raíces cuadradas o resolver ecuaciones, está emulando la inteligencia humana, porque es evidente que para esos cálculos se precisa un poco o mucha inteligencia. De hecho, la mayoría de la población del norte global con estudios de secundaria tiene problemas para realizar determinados cálculos sin calculadora. Sin embargo, nadie se siente amenazado por las calculadoras. La inteligencia de las calculadoras no suscita inquietudes.

Un asistente de trayectoria que corrige los desvíos del vehículo dentro del carril, sistema ya integrado en muchos de los coches que circulan por las carreteras, también emula la inteligencia humana que se necesita para conducir sin salirse de la vía. De hecho, atendiendo a cómo está construido, es inteligencia artificial. Pero, de nuevo, nadie se siente amenazado por ese automatismo. Su inteligencia no resulta preocupante.

Poner como criterio la inteligencia humana abre un cajón en el que cabe todo. Es un lío meter en el mismo saco los cálculos

matemáticos, el reconocimiento de imágenes, la práctica del ajedrez, el asistente de trayectoria, el diagnóstico médico, tirar bombas, la traducción profesional, la prevención de la violencia de género o el cuidado de las ancianas.

Aunque la definición popular dice que toma como referencia la emulación de la inteligencia humana, en realidad no lo hace así, sino que utiliza arquetipos que se han construido con narrativas, con celofanes, con ciencia ficción o con propaganda, mezclas provenientes de diferentes fuentes. En inteligencia artificial, los principales arquetipos son: la inteligencia artificial conversadora, que maneja el lenguaje natural —como Siri, Alexa o ChatGPT—; la androide, dotada de cuerpo —como el nietecito robótico de Corea del Sur—; la solucionadora de problemas, que resuelve tareas —como AlphaZero o Apertium—; y la predictiva, que modela un ámbito de la realidad —como Lavender o VioGén—. Lo que queda fuera de estos arquetipos no se considera intuitivamente inteligencia artificial.

Así que por una parte tendríamos un saco muy genérico en el que cabe todo y por otra unos arquetipos muy concretos cuyo imaginario deja fuera, por ejemplo, todo lo relativo a la gestión de infraestructuras críticas, como el tráfico por carretera, el suministro de agua, gas o electricidad, la predicción atmosférica, los antivirus y otros cortafuegos para protegerse de ciberataques o el diseño de vacunas.

Pero el caso es que la inteligencia artificial existe. Es algo. Para regular ese algo, la Unión Europea ha promulgado una reglamento que entrará en pleno vigor en 2026. En él excluye los sistemas basados en las normas definidas únicamente por personas físicas para ejecutar automáticamente operaciones (es decir, las calculadoras) y establece que la característica principal de los sistemas de inteligencia artificial es su capacidad de inferencia. La inferencia es un proceso por el cual se derivan conclusiones a partir de premisas o hipótesis iniciales. Dicho de otra manera, consiste en utilizar información para procesarla con el fin de emplearla de una manera nueva o diferente (como hace el asistente de trayectoria). Cuando ese proceso lo realiza un humano, las inferencias son el proceso mental que extrae significado o conclusiones a partir de pistas, indicios o datos indirectos.

Para el reglamento europeo, «esta capacidad de inferencia se refiere al proceso de obtención de resultados de salida, como predicciones, contenidos, recomendaciones o decisiones, que puede influir en entornos físicos y virtuales, y a la capacidad de los sistemas de IA

para deducir modelos o algoritmos, o ambos, a partir de información de entrada o datos. Las técnicas que permiten la inferencia al construir un sistema de IA incluyen estrategias de aprendizaje automático que aprenden de los datos cómo alcanzar determinados objetivos y estrategias basadas en la lógica y el conocimiento que infieren a partir de conocimientos codificados o de una representación simbólica de la tarea que debe resolverse. La capacidad de inferencia de un sistema de IA trasciende el tratamiento básico de datos, al permitir el aprendizaje, el razonamiento o la modelización. […] Los sistemas de IA están diseñados para funcionar con distintos niveles de autonomía, lo que significa que pueden actuar con cierto grado de independencia con respecto a la actuación humana y tienen ciertas capacidades para funcionar sin intervención humana». Incluyen estrategias de aprendizaje automático, pero se sobreentiende que también incluyen otras que no lo son.

¿Y por qué hay que regular estos sistemas? Porque su uso entraña riesgos. Siguiendo el reglamento europeo, riesgos para los intereses comunes tales como la seguridad física de las personas, la salud, los derechos y libertades fundamentales, la democracia, el Estado de derecho o la protección al medio ambiente, y que se clasifican como inaceptables, altos, específicos o mínimos. El reglamento es un intento de mitigar esos riesgos. Los movimientos por la justicia algorítmica demandan una implementación estricta y fuerte del reglamento; medidas urgentes para resolver los vacíos legales —especialmente los que afectan a las personas migrantes—; prohibición explícita para el reconocimiento biométrico a distancia y el reconocimiento de emociones en todas sus formas; y definición de estándares para la transparencia y la inclusividad.

Pero, más allá de lo que diga el reglamento, desde un punto de vista tecnosocial, en estas máquinas computacionales que hacen inferencias, ¿qué es lo amenazante y dónde está?

Los bebé *reborn* son muñecas artesanales hechas con vinilo o silicona. Tienen unas características tan realistas que pasarían por ser un bebé humano. Cada bebé es único. Tienen venitas, rojeces, lunares… Las pestañas y el cabello se injertan pelo a pelo. El peso está calculado para que al cogerlos en brazos dé la sensación de estar cogiendo a un bebé real. Cuestan desde cientos hasta miles de euros. No se compran, se adoptan.

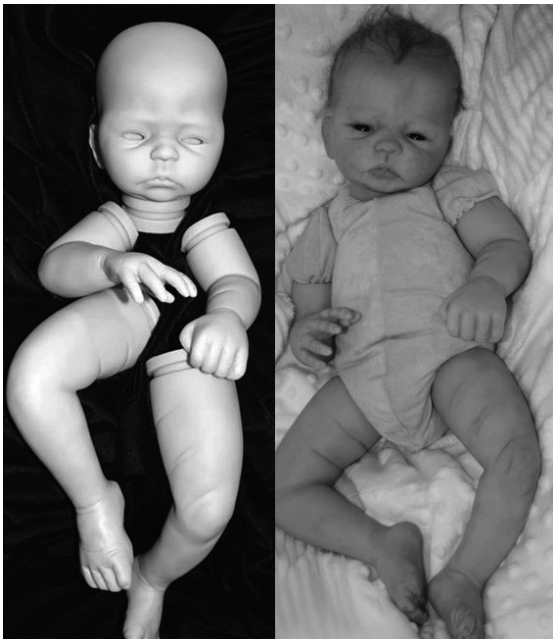


Imagen 30. El realismo de estos bebés *reborn* es impactante, pues verdaderamente parecen humanos. Fuente: https://es.wikipedia.org/wiki/Beb%C3%A9_reborn#/media/Archivo:Donna-Lee-Emmaline-comparison.jpg

Un bebé que no crece, no llora, no ensucia, no enferma, no da malas noches, pero al que puedes hablar, pasear, bañar, peinar, vestir, llevar en brazos o dormir. Proporcionan paz y calman la angustia cuando se atraviesa un mal momento, una depresión, un Alzheimer, o cuando el dolor por la pérdida de un hijo humano se hace insoportable.

Abierto en 2024 en Berlín, Cybrothel es un burdel para practicar sexo con muñecas sexuales de tamaño real que cobran vida en un entorno de realidad virtual. Ofrece *suites* privadas y acompañantes con nombres seductores, solo que los clientes, en lugar de relacionarse con trabajadoras sexuales humanas, lo hacen con muñecas sexuales de tamaño real dotadas con inteligencia artificial. Las muñecas no pueden hablar ni moverse, pero con gafas de realidad virtual se puede vivir una experiencia de pornografía o de sexo virtual y también se puede hacer *sexting*. Los promotores hablan de que su negocio consiste en ofrecer un espacio seguro en el que los clientes pueden disfrutar

sin angustia, al tiempo que señalan que en Alemania, después de la pandemia, la industria del sexo ha visto una disminución significativa en la cantidad de trabajadoras sexuales disponibles. Cuesta creer este último dato.

Con la inteligencia artificial no solo entra tecnología. Con ella también entra la ideología que legitima su penetración: las bondades de una vida conducida por algoritmos.

La ideología de la inteligencia artificial dice que tu vida será mejor si, en lugar de rodearte de personas como tú, te rodeas de asistentes a los que puedes dar órdenes y endosarles los trabajos sucios, los trabajos que pueden hacer por ti, para que tú te entregues a otras actividades de más alto rango. Señala, así, una jerarquía: tareas que mereces hacer, tareas que no mereces hacer.

También dice que la subjetividad es mala: una debilidad humana a superar, para alcanzar la objetividad de un mundo que está ahí fuera y tiene sentido e independencia por sí mismo, al margen del sesgo humano. Por eso, las decisiones humanas deben estar legitimadas por máquinas. La imperfección es algo a eliminar y las máquinas pueden hacerlo.

Y que está bien poder configurar a tu gusto a las personas que te rodean, y tienes derecho a hacerlo. Si no responden a tu configuración, si te contradicen o te perturban, te niegan, te cuestionan, te interrumpen, te molestan…, tienes derecho a barrerlas de tu horizonte para poder hacer tu camino libre de otros que te planteen sus propios deseos, necesidades o neurosis, que casi nunca estarán alineados con los tuyos. Si nadie te da la conversación que necesitas, ChatGPT siempre estará ahí.

El programa de la inteligencia artificial, su celofán, es la creencia de que un mundo sin subjetividad es un mundo mejor. No más cansancio, no más disputas, no más soledad, no más errores, no más imperfecciones, no más enfermedad, no más muerte… ¿No más humanidad?

La inteligencia artificial, junto con la robótica y la biogenética, es uno de los pilares del movimiento transhumanista: una crítica de la comprensión actual de lo humano que pone el acento en las posibilidades de un perfeccionamiento que venza a la enfermedad y dé muerte a la angustia por la muerte y a la muerte misma. Los gurús de este movimiento viven en Silicon Valley.

Elon Musk cofundó la empresa Neuralink, que hace ensayos para implantar chips en cerebros humanos con fines terapéuticos, para que personas con lesiones en la médula espinal puedan volver a caminar. Muy bonito, pero su objetivo último es conseguir humanos radicalmente mejorados, mejorados hasta el infinito, hasta la vida eterna. Es su proyecto transhumanista.

Se dirá que la mayoría de las personas razonables consideran que el transhumanismo es una ida de olla futurista sin ninguna base. Cierto, pero no tanto.

Almudena Hernando explica que el humano, con su inteligencia, comprende la inmensidad y la complejidad del universo y siente angustia. En su deseo de aplacar esa angustia, en su búsqueda de seguridad, poco a poco va virando hacia la tecnología. Se abre paso primero la especialización, dentro del grupo, y luego las relaciones de poder. A partir de un cierto umbral antropológico o histórico, individuos dentro de los grupos humanos empiezan a valorar que disponer de tecnología da más seguridad que la que proporciona la pertenencia al grupo. Se trata de un viraje masculinizado, pues la socialización de las mujeres sigue ligándolas a lo grupal y lo comunitario.

Al cabo de miles de años, el uso de la tecnología continúa aportando sensación de control, de dominio sobre la realidad, reduce la incertidumbre, calma la ansiedad. Mejor tecnología que vínculos sociales. Google Maps da más seguridad que preguntar a otros humanos del lugar. Una videoconferencia da más seguridad que un encuentro cara a cara. ¿Transhumanismo a pequeña escala?

El transhumanismo es un movimiento amplio que engloba distintas corrientes, algunas de ellas opuestas entre sí. En su versión tecnocientífica propone un mejoramiento humano mediante la integración humano-máquina o mediante la manipulación genética. Asume que nuestro soporte biológico material es defectuoso, así que sus líneas de trabajo son: mejorar la salud modificando genes relacionados con enfermedades o incluso sustituyendo el genoma entero para conseguir inmunidad ante todos los virus; alargar la vida y, tal vez, conseguir la inmortalidad, pues la vejez se considera una enfermedad; aumentar la inteligencia o verter la mente de un humano en una máquina para hacerlo inmortal, y mejorar la memoria con implantes y cosas por el estilo.

La idea fuerte detrás de esta propuesta es: si tenemos la oportunidad de eliminar enfermedades o de superar limitaciones, evitando

así sufrimiento, ¿por qué deberíamos no hacerlo? Es más: ¡tenemos la obligación moral de hacerlo! El argumento viene a decir que la evolución al estilo Darwin es muy lenta y además es indeterminada, pues no tiene una finalidad y nada garantiza que la evolución no nos haga involucionar o desaparecer. Así que ha llegado el momento de que la especie humana se descuelgue de esa evolución y tome un atajo para evolucionar, es decir, mejorar, según sus propios objetivos. La humanidad puede y debe tomar las riendas de su propia evolución. Por ejemplo, autoaplicarse mejoras que le permitan resistir al cambio climático. En la versión más audaz se habla de que ese conjunto de mejoras, ese humano mejorado, daría lugar a una nueva especie, que ya han bautizado como *Homo excelsior*. En algunas variantes se propone una modificación genética para eliminar la agresividad humana y alcanzar así ese mundo mejor que ninguna de las utopías ha conseguido construir. En otras se dice que, dado que dentro de poco va a haber robots mucho más inteligentes que los humanos, si la humanidad quiere tener chances, debe mejorarse a sí misma o, de lo contrario, quedará condenada a ser esclava de los robots.

Puede parecer ciencia ficción, pero no lo es. Esas líneas de trabajo están ya en los laboratorios y en la industria. Como dice Natasha Vita-More, directora de Humanity+, una de las asociaciones pro-transhumanismo más activas: «Piense como un futurista en la industria de la longevidad de cien mil millones de dólares».

Junto al transhumanismo tecnocientífico se sitúa un transhumanismo político, que a veces se denomina poshumanismo, sobre el que existe la controversia de si es transhumanismo o no. Se trata de una corriente de pensamiento que tiene su origen en los años en los que se desarrolló la cibernética. La cibernética modificó radicalmente las ideas mediante las que se explicaba la identidad humana, que quedó despojada de su singularidad y pasó a ser un elemento más dentro de un mundo de sistemas acoplados organizado por flujos de información. La cibernética descorporeizó la razón. La razón ya no residía exclusivamente dentro del cuerpo humano, también podía existir dentro de una máquina inteligente. Lo humano reconocía en lo no humano a un igual y en adelante se relacionaría horizontalmente con ello.

El poshumanismo coloca lo humano en una posición de horizontalidad respecto a la naturaleza no humana y a los agentes maquínicos, y propone una convivencia igualitaria entre todas las bandas, un ensamblaje colectivo y comunitario sin jerarquías entre humanos y no humanos. Es una idea filosófica y política retomada

por el posmodernismo y desarrollada, entre otras, por Donna Haraway en su conocido *Manifiesto cíborg*.

Mientras que el transhumanismo tecnológico considera que la humanidad está en un estado embrionario que, mediante la aplicación intensiva de tecnologías y de bioingenierías, tiene un amplio recorrido de mejora y vislumbra una aceleración evolutiva que daría lugar a una transhumanidad perfeccionada, este transhumanismo político critica un universo centrado en lo humano y la arrogante creencia en nuestra singularidad y superioridad.

Propuestas que trastabillan la identidad humana y que permiten lecturas tanto desde la izquierda como desde la derecha, pero con una idea común: disolver las fronteras entre humano y máquina es una liberación. La gran diferencia entre ellas es la respuesta a la pregunta sobre de qué hay que liberarse, pues mientras que para Donna Haraway hay que liberarse de las identidades sobre las que se instalan las relaciones de poder, para Natasha Vita-More hay que liberarse de las limitaciones físicas a la longevidad. Haraway está viendo relaciones sociales, relaciones de poder, conflictos y opresión. Vita-More está viendo humanos aislados, una naturaleza que está ahí para ser cambiada a voluntad y un mercado por conquistar.

Pese a estas grandes diferencias, pese a todas sus discrepancias de fondo, ambas propuestas comparten un corolario común: no hay una naturaleza humana que deba ser preservada.

Ciertamente, no es una tontería. Es un llamamiento a replantear el sentido de lo humano.

Entender

Entender la tecnología no es sencillo, porque está muy presente y a la vez está muy oculta. La tecnología dispara creencias, ideas *a priori*. Y una de ellas consiste en creer que es neutra y que su bondad o maldad depende del uso que se le dé. Un piolet puede servir para caminar con seguridad en pendientes con hielo o para asesinar a Leon Trotski. Todo depende de cómo se use.

Pero es muy raro que una actividad humana, como la producción de tecnológicas, que tiene lugar en sociedades con sistemas económicos, ideologías y relaciones de poder concretos, sea neutral. Es muy raro que las tecnologías no queden impregnadas de los valores con los que han sido diseñadas.

Tomemos por ejemplo el caso de la tecnología coche. Se dirá que un coche se puede utilizar para llevar comida al banco de alimentos (uso bueno) o para atropellar a tu pareja y librarte de ella (uso malo). El coche sería neutral, ni bueno ni malo. Todo depende de cómo se usa.

Pero un coche suelto, como objeto aislado, no es nada. Los coches son una pieza en una malla de conexiones que incluye las fábricas en las que se construyen, las redes de concesionarios que los venden, la fabricación y distribución de piezas de recambio, la cercanía de talleres donde se reparan, la circulación del conocimiento y la mano de obra disponible para esas reparaciones, la regulación de las normas de circulación, el sistema de inspecciones técnicas, las autoescuelas donde se aprende a conducir, la disponibilidad de combustible y estaciones de servicio que lo comercializan, la red de calles con pavimentos adecuados, las carreteras con curvas lo suficientemente abiertas para que circulen a velocidad alta o media

con eficacia… Sin todo ello, el coche sería un objeto de museo, algo inservible más allá de ser portador de memoria histórica o de evocaciones románticas, como lo pueden ser ahora una máquina de escribir, un fax, una cabina telefónica o una enciclopedia en tomos. Una vez se ha desmontado la malla de conexiones en las que la máquina de escribir, el fax, la cabina o la enciclopedia cobraban sentido, ya no son más que antiguallas de las que cuesta deshacerse, pero que son inservibles.

Viendo el sistema completo, es fácil observar cómo el automóvil, desde su invención a finales del siglo XIX, ha impactado en la movilidad de las personas y en el urbanismo de las ciudades. Cómo ha transformado la vida en común.

El coche precisa de infraestructuras colectivas que se financian públicamente. Consume recursos y energías. Incluso si va completamente ocupado, es el medio de transporte que más energía consume por persona transportada y kilómetro recorrido. Genera residuos. Es el principal emisor de la contaminación del aire en las ciudades. El automóvil es responsable del 80 % de emisiones de NO_2 debidas al tráfico y del 60 % de emisiones de partículas. Ocupa espacio urbano común —la superficie urbana dedicada al automóvil es de entre el 20 % y el 30 % del total, y en urbanizaciones de nueva construcción alcanza porcentajes del 40 %—. Produce atascos y ruidos —el 80 % del ruido urbano se debe al tráfico rodado—. A todo ello hay que añadir que es el medio de transporte que causa más muertes y heridas. Es el principal causante de lesiones y de muerte entre las personas jóvenes. Y una trampa mortal en caso de riada.

¿Cómo es posible que una sociedad acepte de buen grado el uso de una tecnología contaminante, ruidosa, que se adueña del espacio público y cuyos accidentes mortales superan anualmente el millón de personas en el mundo? Pues es posible porque el coche es portador del valor de la libertad individual: libertad de ir a donde se quiera, cuando se quiera, por donde se quiera y con quien se quiera. El automóvil se convirtió en símbolo de prosperidad y libertad. Una libertad que, en aras de la democracia, debía ser disfrutada por todas las personas. No era justo que esa libertad la disfrutaran solo unos pocos privilegiados. De ahí el programa de la fábrica de coches Ford: el Model T, un coche barato y accesible para todos. El primer vehículo producido en cadena, fabricado desde 1908 hasta 1927. El primer coche para el mercado de masas.

Un coche para las masas es algo muy concreto que penetra en la sociedad de una manera muy concreta. Una ambulancia, un coche de bomberos y un coche familiar, todos tienen un motor de explosión. Tienen una tecnología común, pero de nuevo un motor de explosión suelto, puesto encima de la mesa, no sirve para nada. Para que sea útil, se tiene que ensamblar con otros componentes, como frenos, sistemas de tracción y también carrocerías o asientos. En sí mismo es una malla de conexiones, de ensamblajes entre tecnologías dispuestas y conectadas entre sí de un modo muy preciso y determinado: sus dos asientos independientes delante para el matrimonio; su banco corrido detrás, para que se puedan apiñar la abuela y los niños; su maletero; sus ventanillas; sus dimensiones; sus colores…

Una ambulancia o un coche de bomberos destilan un halo de servicio o incluso de heroicidad, pero no de libertad. En cambio el coche sí, y esta identificación entre el coche y la libertad individual como valor, con la consecuencia lógica de su democratización, es lo que ha dado lugar al sistema de movilidad que tenemos: cada familia que puede permitírselo tiene uno o más vehículos. Ir en coche es un derecho y limitar su uso merma las libertades. Con otro sistema de valores, los coches podrían ser un recurso limitado y comunitario que se utilizara solamente en casos excepcionales de trayectos muy específicos en los que no haya medio de transporte colectivo, o para personas con dificultades para la movilidad, o que se alquilara puntualmente por el placer de viajar. En su justa medida, podría ser una tecnología muy conveniente que facilitase la vida.

Pero, claro, para ello habría que cambiar muchas cosas. Habría que desmantelar un inmenso mercado, fortalecer otras mallas de conexión, como los transportes públicos y comunitarios, y también desplazar el valor de la libertad hacia otros imaginarios: un planeta libre que respira aire puro, unas criaturas libres que juegan en la calle sin atropellos, un espacio público vacío y libre propicio para el encuentro… En cambio, los valores con los que el coche ha sido diseñado, su celofán, son el éxito y la libertad personales. No tener coche es de pobretones. La supremacía del coche es natural.

Cuesta mucho desplazar una palabra de prestigio. Las palabras de prestigio se resisten a ser relegadas. No quieren ir a menos. Y «libertad» tiene prestigio. ¿Quién me va a prohibir a mí ir con mi coche por donde me dé la gana? Pero incluso cuando hay conciencia social y ecológica, incluso cuando su uso es consciente y para fines buenos,

incluso si tienes coche y solo lo usas porque así puedes ir a atender a tu madre cuando sales del trabajo, el coche no es neutro. Sigue contaminando, ocupando el espacio público, produciendo atascos y ruido. Otra cuestión es qué hacer al respecto.

Las limitaciones a la circulación que se aplican en los centros urbanos, el aumento de los precios de los carburantes o la aplicación de impuestos sobre el carbono está dando lugar a luchas como la de los chalecos amarillos, protagonizadas por personas que viven en la periferia y tienen coches viejos y vidas precarias. Luchas a las que cuesta aplicar el esquema de izquierdas y derechas, y que señalan que esa medida supuestamente ecologista consistente en remover solo una pieza de la malla sustituyendo el coche de gasolina por el eléctrico es un lujo que no se pueden permitir.

Cuando se dice que una tecnología es neutra, es porque no se están comprendiendo sus efectos. Y el principal efecto, a lo que más hay que atender, es a cómo transforma las relaciones sociales. Para poder ver en qué sentido una tecnología no es neutral, hay que mirar desde arriba, contemplar toda la trama y observar cómo esa trama cambia la realidad social. A principios del siglo XX la producción de coches en cadenas de montaje, dirigida al mercado de masas, transformó la movilidad y reconfiguró las ciudades. Está por ver cómo se va a reconfigurar ahora la reacción ante los desastres que el uso desenfrenado de esa tecnología, que tantos beneficios económicos ha generado a la industria automovilística, está produciendo. Está por ver qué exclusiones, qué brechas, qué injusticias, qué consecuencias tiene para las personas pobres. Para el mercado de la automoción, cambiar una pieza (gasolina por electricidad) es más fácil que desmontar toda la trama para construir otro modelo global de transporte que sea radicalmente mejor. La vieja estrategia de cambiar algo para que todo siga igual.

Pero las tecnologías no son algo que haya que tomar tal cual y ya está. Hay posibilidad, hay capacidad para actuar, para entramarlas en mallas de conexiones que protejan a la sociedad contra los malos usos, que mitiguen los efectos de los usos contrarios al bien común.

En la Unión Soviética la industria automovilística se orientó más hacia camiones, tractores y autobuses que hacia coches urbanos de uso personal. La malla soviética fue distinta, posiblemente porque el valor de libertad individual no cabía en sus esquemas.

En 1989 se creó en el Estado español la Organización Nacional de Trasplantes. España cuenta, con diferencia, con el índice más elevado de donaciones de órganos del mundo, hasta el punto de que la literatura científica habla del «modelo español». Se podrá decir que la tecnología del trasplante es neutra. Depende de si se secuestra a un niño para matarlo, extraerle un órgano y trasplantárselo a un rico (uso malo) o si se dona médula ósea para curar el cáncer de un familiar (uso bueno). Pero para ver con más perspectiva hay que elevar el nivel de análisis y observar qué efectos tiene el sistema de trasplantes de órganos en su conjunto. Escudriñar con qué valores está diseñado.

Los trasplantes no se dan de forma aislada, cada uno a su manera. Para hacerlos se precisa de todo un sistema social y sanitario. Un sistema orquestado en tiempo real, ya que la donación y el trasplante no se pueden programar con antelación, pues se producen de manera inesperada y tienen carácter urgente. Un sistema cohesionado para que los centros sanitarios puedan hacer extracciones con eficiencia, para que haya una regulación y las donaciones se hagan con garantías, para que el receptor dé un consentimiento informado, para que los órganos se conserven y se distribuyan según criterios equitativos y sin picaresca… El sistema español se ha construido bajo los valores de lo público, el acceso universal, la reciprocidad entre comunidades autónomas y la no mercantilización. La donación no se remunera. El receptor no paga. Un sistema que se reconoce como el mejor del mundo y del que cabría pensar cuánto de su éxito se debe a la pericia técnica de las profesionales de la medicina y cuánto a los valores con los que ha sido diseñado, o cómo un aspecto realimenta al otro, pues el propio sistema, de por sí, por su diseño, incentiva las donaciones altruistas. Un sistema que no es neutral.

Los protocolos básicos de Internet se desarrollaron bajo el principio de que distribuido es mejor que centralizado. Esta decisión era una valoración. Destilaba valores. Las consecuencias de ello son ampliamente conocidas.

El lenguaje de programación BASIC se desarrolló para que personas no expertas pudieran programar. No era un lenguaje de programación muy bueno, pero era sencillo y abrió la puerta de la programación a muchísima gente. Se diseñó bajo el valor de la accesibilidad.

Cuando se dice que la tecnología es neutra, es porque se está pensando en el producto tecnológico aislado (el piolet, el motor de explosión, el coche, el trasplante). Se está pensando en el resultado.

Pero nada de ello existiría si antes no hubiera habido investigación, es decir, tecnología como proceso. Para investigar hacen falta financiación, recursos… La financiación va para lo que alguien decide que es relevante. Las investigaciones se evalúan, los financiadores exigen resultados. El sistema científico está continuamente conflictuado por injerencias económicas y políticas, pese a las cuales las investigadoras, con sus deseos, con sus anhelos, siguen teniendo agencia. Cuando Marie Curie investigaba sobre la radioactividad en un hangar, contra viento y marea, lo hacía porque quería investigar justamente eso. No le daba lo mismo investigar cualquier otra cosa. Una investigadora tiene ideas, montones de ideas. Tiene aspiraciones, intuiciones, imagina, proyecta, sueña… Se implica.

«Libertad» es una palabra de prestigio. «Inteligencia» también lo es. La tecnología entra envuelta en su celofán de valores. Va a ser muy difícil desplazar el término «inteligencia artificial» por algún otro con menos celofán, tipo «sistema que genera resultados a partir de inferencias sobre los datos de entrada».

Entonces ¿es posible desarrollar inteligencia artificial que no reproduzca lógicas de opresión? Esta es la pregunta que se aborda en el documento *Hacia un marco feminista para el desarrollo de* IA*: de los principios a la práctica*, fruto del trabajo de la red f<A+i>r, que en América Latina trabaja para que la inteligencia artificial sea integradora y transformadora, y no solo eficiente. El documento rastrea relaciones de poder, colonialismo, racismo, clasismo y otros sistemas estructurales injustos, y analiza las experiencias, en el día a día, de siete mujeres que trabajan en algún campo de la inteligencia artificial o la ciencia de datos en la región, en diálogo con distintas declaraciones de principios y guías feministas para el desarrollo y despliegue de tecnologías digitales.

En el documento se afirma que la inteligencia artificial es mucho más que un asunto de cómputo, máquinas y datos. Es fundamentalmente política porque está siendo permanentemente moldeada por un conjunto de prácticas técnicas y sociales, así como de infraestructuras, instituciones y normas. También es material, porque está compuesta por recursos naturales, energía y trabajo humano. Es a la vez una tecnología, una ciencia, un negocio, un sistema de conocimiento, un conjunto de narrativas, de relaciones y un imaginario. Un imaginario en disputa.

Explica que la imaginación ha sido una potente herramienta de disputa frente a los modelos de desarrollo tecnológico impuestos. Como dice Donna Haraway en su *Manifiesto cíborg*, «la liberación se basa en la construcción de conciencia, de aprehensión imaginativa de la opresión, y también de la posibilidad».

La mayor parte de estas tecnologías se desarrollan en Estados Unidos y, que sepamos, por hombres. Y para que no solo les beneficie a ellos es necesario que un grupo diverso de personas se involucre en el desarrollo. La inteligencia artificial tiene que servir a la humanidad, no al mercado. No existe ni jamás existirá una inteligencia artificial que por sí sola produzca transformaciones concretas hacia una sociedad más justa. Sin embargo, es posible cambiar las prácticas de quienes participan en su diseño, producción, despliegue y gobernanza para mitigar sus impactos perjudiciales, como parte de un proceso de transformación hacia una vida social más justa. Es preciso mantener suficiente presión feminista en el desarrollo de las tecnologías, en un contexto donde las consecuencias materiales menoscaban cualquier posibilidad liberadora.

Todo despliegue tecnológico ocurre en un espacio de tensiones políticas y es iluso suponer que las máquinas, por su propia eficacia, van a obviar, ocultar o estar por encima de esas tensiones. En lugar de preguntarnos cómo desarrollar y desplegar un sistema de inteligencia artificial, ¿no deberíamos preguntarnos primero por qué construirlo, si es realmente necesario, quién lo pide, quién gana y quién pierde?

Pensar

Carta a Santi

Querido Santi, ¿cómo estáis?

Ya te he enviado los capítulos sobre filosofía, por si ves que hay que corregir algo. Me siento un poco insegura con la filosofía. Bueno, y con todo lo demás. Ja, ja, ja.

Voy a empezar por Marx, porque eso de la propiedad de los medios de producción parece una cosa desfasada, pero creo que todavía tiene su aquel. El fragmento de las máquinas sigue siendo lectura obligada :-)

En el libro de *Riders on the Storm*, Núria explica cómo, en la lucha de los **riders**, para que se les reconociera la relación laboral con las plataformas de *delivery*, establecer de quién era la app que les permitía trabajar fue crucial. Se puede ser repartidor sin moto, sin bicicleta…, pero no se puede repartir sin la app. Y la app es de la plataforma. La app es como el medio de producción, que es del amo. Y eso fue el punto fuerte en la demostración de la relación laboral, lo que les permitió luchar por la idea de que los riders no son trabajadores autónomos, sino asalariados, y llevarlo a los tribunales.

Bueno, seguiré por Heidegger. Su pasado nazi es un poco…, ejem…, pero al fin y al cabo es uno de los pensadores contemporáneos más influyentes y siempre dices que ha pensado de un modo muy radical qué significa pensar, ¿no?

Y terminaré por Simondon. Deleuze lo leía y un día me explicaste que el concepto de singularidades preindividuales lo sacó de cómo Simondon pensó la individuación. No te creas que lo entiendo del todo, pero la gracia está en pensar lo que no se entiende :-)

Toni me ha dicho que ponga más filósofos de la ciencia, pero al final me quedo con estos tres. Es que no quiero que el libro sea la historia interminable.

También podría poner ideas cruzadas, porque al final unos pensadores conversan con otros. Simondon no veía la cibernética como Wiener. Es un filósofo a la vieja usanza, pero plenamente inspirado en la cibernética y en las teorías de la información. ¡Un crack! Pero no me veo capaz de elaborar esos cruces.

Y he dudado si poner a Miguel Benasayag. Me he leído el de *Funcionamos o existimos*, que lo plantea como una respuesta a la colonización algorítmica. El de *La inteligencia artificial no piensa todavía* no lo he leído. Me figuro que estará muy bien, pero me rompe un poco el flujo, porque pasar de Simondon a las visiones hacker queda en continuidad y, si pongo algo en medio, pues… ya no me cuadra tanto.

Bueno, Santi, ya me dirás. A ver si me puedo apuntar a la próxima cena, me cuentas también sobre tu libro y nos ponemos al día :-) Muchas ganas de leerlo.

Un fuerte abrazo.

6. Filosofía

Marx

En Europa, 1848 fue un año de revoluciones. Levantamientos protagonizados por una clase obrera que a partir de sus derrotas entendió que, como movimiento obrero autónomo, debía autoorganizarse en sindicatos y partidos políticos propios para luchar por sus intereses de clase proletarios. También fue el año en el que Karl Marx y Friedrich Engels publicaron el *Manifiesto comunista.*

Por esas fechas, desde el pensamiento filosófico, desde las observaciones del funcionamiento de los motores y máquinas y desde los experimentos de laboratorio, se estaba estableciendo el principio físico de conservación de la energía: la naturaleza en su conjunto posee una reserva de fuerza (energía) que no puede de ninguna manera ser aumentada ni disminuida. Por tanto, la cantidad de fuerza en la naturaleza es igual de eterna e inalterable que la cantidad de materia. La energía no se crea ni se destruye, solo se transforma.

De alguna manera, todas las energías son una misma fuerza que, en la naturaleza, se organizan según distintos niveles de complejidad. La electricidad se puede transformar en magnetismo, la energía mecánica se puede convertir en eléctrica, el calor se puede transformar en movimiento, la electricidad se puede transformar en calor… Quedaba así establecido un principio universal de enormes consecuencias prácticas para la construcción de las máquinas en plena Revolución Industrial. Una máquina (o motor) que realice un trabajo físico necesita una fuente de energía y no puede suministrar o convertir en trabajo más energía que la que obtiene de la fuente. Las máquinas no pueden crear ni destruir energía, solo pueden transformarla.

Mientras en la ciencia física se establecía el concepto de «trabajo» y se definía su formulación matemática, el capitalismo estaba industrializando la sociedad, concentrando las fábricas en ciudades y provocando las migraciones y la proletarización de millones de personas que se veían forzadas a dejar atrás sus formas de vida tradicionales y artesanales para convertirse en obreros asalariados. Los trabajadores proletarios, hombres y mujeres, niños y niñas, trabajaban sin descanso dieciséis horas al día a cambio de un mísero salario y vivían en infraviviendas en ciudades sin agua corriente ni alcantarillado y, por supuesto, sin acceso a la educación, a la salud ni a la cultura. Pero no eran víctimas pasivas. Estaban tomando conciencia de clase, conciencia de pertenecer fraternalmente a la misma clase social explotada por la burguesía. Eran la clase obrera y tenían conciencia de clase, conciencia de tener intereses comunes compartidos y de que con la lucha y la solidaridad podían destruir las relaciones de explotación capitalistas y alumbrar una nueva sociedad con relaciones humanas libres entre iguales: la sociedad sin clases.

Karl Marx, en conversaciones con su amigo Friedrich Engels, observaba atentamente tanto las luchas y la toma de conciencia de la clase obrera como el desarrollo de la ciencia de la energía y del maquinismo. Estudiaba cómo el sistema técnico y la organización social convergían en un sistema de producción: el capitalismo. Y se preguntaba: ¿qué es una máquina? ¿Por qué las fábricas se maquinizan? ¿Y qué efectos produce la mecanización en los obreros?

Según su punto de vista, una máquina es muy diferente de una herramienta. La herramienta, el martillo, es un instrumento que transmite la actividad del obrero al objeto. Pero, mientras que la herramienta es un objeto técnico inerte que debe ser animado por la destreza y la intención del obrero, la máquina, conectada a una fuente de energía, a un motor, toma el lugar del obrero. La máquina modifica directamente la materia prima, mientras el obrero se limita a vigilar su funcionamiento. Ella misma tiene destrezas, a causa de las leyes mecánicas que operan en su interior. No es un medio de trabajo para el obrero. Ocupa su lugar, porque cristaliza, en su funcionamiento, la habilidad equivalente a manipular un conjunto amplio de herramientas con orientación a un fin.

Marx distingue entre fuerza y trabajo. La fuerza es una capacidad que poseen por igual entidades naturales —como el agua o el viento—, entidades animales —como los bueyes o los humanos— y entidades artificiales —como los motores de vapor—. En cambio,

el trabajo es la capacidad de modificar intencionalmente la materia con un propósito decidido, en el que cabe cierto margen de libertad. El trabajo es una facultad exclusivamente humana, producto de las habilidades físicas y mentales de los seres humanos. Así lo explica: «Concebimos el trabajo bajo una forma en la cual pertenece exclusivamente al hombre. Una araña ejecuta operaciones que recuerdan las del tejedor y una abeja avergonzaría, por la construcción de las celdillas de su panal, a más de un maestro albañil. Pero lo que distingue ventajosamente al peor maestro albañil de la mejor abeja es que el primero ha modelado la celdilla en su cabeza antes de construirla en la cera. Al consumarse el proceso de trabajo surge un resultado que antes del comienzo de aquel ya existía en la imaginación del obrero, o sea idealmente».

¿El trabajo pertenece exclusivamente al hombre?, se pregunta Marx. Y su respuesta contundente es que sí, que la capacidad de trabajo es un atributo exclusivamente humano. Sin embargo, con la mecanización aparecen máquinas que transforman la energía para producir mercancías. No trabajan, pero producen. Esto es inédito en el mundo, es una revolución. Es la Revolución Industrial.

En la génesis de la máquina, ciencia y capital se encuentran. La maquinaria industrial es un sistema automático conectado a un motor, otra máquina, que ya no puede ser accionado por energía humana o animal, pues la fuerza necesaria para vencer las resistencias físicas de ese automatismo maquínico supera las capacidades del cuerpo humano. Las maquinarias se encadenan y se organizan en cadenas de producción, unas conectadas a las otras, creando sistemas automáticos de máquinas. Ahora la actividad del obrero está regulada y determinada en todos los aspectos por el movimiento de la maquinaria, y no a la inversa. El conocimiento del obrero sobre cómo manejar una herramienta ya no tiene valor. Lo que cuenta es el conocimiento científico, capaz de construir máquinas. Pero, en la organización social, el obrero no es un científico. No es poseedor de la ciencia, sino que esta opera a través de la máquina, como poder ajeno, como poder de la máquina misma sobre él.

Marx se detiene a ver qué es lo que ocurre con el trabajo asalariado y con las máquinas en el capitalismo. La riqueza de la vida, la transformación de la naturaleza, lo que podría ser un ejercicio creativo y libre fruto de ensamblajes cooperativos entre humanos y máquinas, en el capitalismo solo puede desarrollarse como producción de mercancías bajo la forma de trabajo asalariado. La producción

de mercancías estresa a las máquinas, les saca todo el jugo, las tiene trabajando veinticuatro horas al día, y hace lo mismo con el obrero, porque ambos son componentes de la misma dinámica productiva: la producción de valor. En las sociedades capitalistas el trabajo toma la forma de «trabajo asalariado», por lo que deja de ser una actividad vital para convertirse en un simple medio de supervivencia. Marx habla, incluso, de una vida expropiada y mutilada. Trabajo a cambio de salario. El capitalista compra energía humana (o sea, la fuerza de trabajo) que da valor a las mercancías, pero, mientras las mercancías adquieren valor gracias el trabajo humano incorporado, el obrero queda reducido a ser una mercancía más inscrita en el mercado laboral. Es la explotación. El trabajo asalariado se gira contra el trabajador, pues es fuente de fatigas, amargura y miseria vital. Un precario medio de subsistencia que empobrece su espíritu. Su trabajo ni le gusta ni le pertenece. Durante la jornada de trabajo el obrero está fuera de sí, no es él, su vitalidad está secuestrada. Solo cuando termina la jornada vuelve a ser él, recobra la alegría, se siente vivo. Es la alienación.

Para Marx, hay una distinción entre el trabajo que está haciendo el obrero en un momento dado, el trabajo en tiempo real, que él llama trabajo vivo, y el trabajo pasado, el que ya se ha hecho con anterioridad, que es el trabajo muerto. Las máquinas son trabajo muerto, son condensaciones de trabajo que se hizo anteriormente y que culminó en la producción de una máquina que, aunque sirve para producir mercancías, ella misma es también una mercancía que el capitalista compró. De estos dos tipos de trabajo, el vivo y el muerto, el único que produce valor (de cambio) es el trabajo vivo. El trabajo muerto no produce valor. Transfiere valor, pero no lo produce. ¿Por qué?

En la actividad que ejecutan las máquinas, Marx observa una especie de ley de conservación de la energía, aplicada al valor: una ley de conservación del valor (de cambio). Si una máquina de fabricar cucharas cuesta 1.000 euros y puede fabricar 1.000 cucharas, después de lo cual se habrá desgastado y ya no funcionará, la máquina transfiere a cada cuchara el valor de 1 euro. Por cada cuchara, la máquina ha producido una cuchara que ha adquirido el valor de 1 euro y ella misma, por su desgaste, vale 1 euro menos.

Sin embargo, el trabajo vivo, el que está realizando el obrero en un momento dado, el que no solo consume fuerza, sino también energía vital, no importa si ese trabajo consiste en apretar el botón que enciende la máquina o consiste en hacer cucharas con sus propias manos, eso sí que crea valor. De hecho, es lo único que crea valor (de

cambio). El trabajo muerto creó valor mientras se hacía, pero una vez hecho ya no crea más. Pero el trabajo vivo sí crea valor y precisamente en eso consiste la explotación del trabajo asalariado. El obrero, con su trabajo asalariado, genera más valor que el que recibe por el salario. Ese plus de valor, esa plusvalía, es la ganancia capitalista.

¿Por qué el trabajo vivo es lo único que crea valor? Porque es fruto de relaciones sociales. A medida que el capitalismo se desarrolla, el proceso de producción se va haciendo cada vez más social, es más dependiente de las relaciones sociales, requiere mucha más cooperación. Mientras que por una parte la producción de mercancías se hace compleja y precisa cada vez más cooperación entre trabajadores con distintas habilidades físicas o intelectuales, con distintos tipos de virtuosismo y de conocimientos, por otra parte el capitalismo tiende a concentrar la propiedad de los medios de producción. Unos pocos capitalistas poseen una gran cantidad de medios de producción que ahora son trabajo muerto, pero que han sido producidos por procesos de cooperación social.

A ese conjunto de conocimientos más o menos concretos o abstractos, de habilidades manuales e intelectuales, de capacidades creativas de todo tipo, de saber experiencial producido colectivamente por la sociedad en su conjunto mediante las relaciones sociales y cristalizado en las máquinas, en las tecnologías y en los procesos productivos como trabajo muerto, a todo ese conjunto Marx lo llama intelecto general *(general intellect).* El intelecto general es una riqueza a la que contribuyen tanto los conocimientos prácticos de la camarera que sabe servir un café como los conocimientos intelectuales del científico que está descifrando el genoma de la planta. Es una riqueza producida socialmente y privatizada, expropiada por los capitalistas, que ejercen la explotación y el mando por medios coercitivos y, en último extremo, represivos.

La vigencia del marxismo en el siglo XXI es una cuestión matizable. Sin embargo, sus principios generales señalan con rotundidad algunas cuestiones.

Una inteligencia artificial es trabajo muerto en tanto que es una máquina más o menos virtual, pero máquina al fin. Desde la perspectiva marxista, no produce valor de cambio. Si dejamos por un momento de lado los costes medioambientales, en cada uno de sus resultados o creaciones produce un pequeño valor de cambio equivalente al valor que ella misma pierde, que es tan pequeño que tiende

a cero. Llevado al extremo, la traducción que hace Google Translator o el texto que genera ChatGPT no tienen valor (de cambio). Lo que sí crea valor es la acción humana de hacer la pregunta o de pegar el texto a traducir. El valor que crea una inteligencia artificial, en cada uno de sus resultados, es insignificante.

Una inteligencia artificial, tanto si es simbólica como si es conexionista, es intelecto general cristalizado. Los algoritmos, y muy especialmente los datos con los que se la entrena, han sido producidos cooperativamente por miles o millones de humanos a partir de sus relaciones sociales. Partidas de ajedrez jugadas, en el caso de Stockfish. Traducciones, en el caso de Apertium. Cientos de miles de millones de palabras, en el caso de ChatGPT, incluyendo tanto contenidos de la web como entradas en blogs personales, artículos en enciclopedias como Wikipedia, textos de revistas y libros, y materiales textuales de todo tipo.

La fundación Wikimedia anda detrás de ChatGPT y otras máquinas generativas similares para ver si han hecho un uso justo socialmente y legal según las distintas legislaciones, al entrenar con datos tomados indiscriminadamente de Internet. Específicamente, la investigación de Wikimedia va en la línea de indagar si se han producido violaciones de la licencia Creative Commons, al no otorgar, en las obras generadas por estas máquinas generativas, la atribución a las fuentes. Las licencias Creative Commons permiten la libre reproducción y reutilización, por lo que programas como ChatGPT podrían copiar el texto de un artículo de Wikipedia o una imagen de Wikimedia Commons, respetando el derecho de atribución de la autoría, es decir, indicando que es una copia y mencionando la autoría. Pero no estamos hablando de un artículo, sino de una copia masiva de contenidos. Está por ver si esa copia masiva puede suponer una violación de la licencia Creative Commons por el hecho de que no se está concediendo la atribución de autoría.

Sin embargo, más allá de que esa copia masiva de obras Creative Commons sea legal o no, desde los movimientos hacker y por la cultura libre se denuncia taxativamente que esta copia masiva rompe el contrato social. No se han creado las licencias libres para eso. No licenciamos las piezas culturales con una licencia Creative Commons para que venga un monopolio, se apropie de todo ello y nos devuelva un producto comercial privativo que nos ofrece adquirir conocimiento obtenido después de haber chupado todo el conocimiento global que anteriormente hemos generado. Es una apropiación del intelecto

general, del cerebro colectivo, que ni reconoce simbólicamente ni, por supuesto, remunera el valor social de los datos que se han utilizado para los entrenamientos. No se trata de algo nuevo en el capitalismo. Todo lo contrario, es capitalismo en estado puro.

Justo en la dirección contraria, la producción de software libre y de datos abiertos, y en general la compartición del conocimiento, sería celebrada por Marx por ser una socialización de los medios de producción, una ruptura de la dinámica capitalista tendente a concentrar la propiedad de los medios de producción en unas pocas manos. El intelecto general *(general intellect)* es perfectamente capaz de diseñar y de socializar inteligencias artificiales de todo tipo y de todo alcance. Si no las produce intensivamente, es porque su acceso al hardware que se requiere para hacerlo con eficacia es limitado. Grandes cantidades de procesadores GPU y TPU, por su precio elevado, solo están al alcance de los capitalistas que han acumulado ingentes cantidades de capital sobre la base de explotar el trabajo humano.

Para Marx no hay nada malo en desarrollar tecnologías, en hacer máquinas. Desarrollar las capacidades productivas es necesario para reducir el tiempo de trabajo, para trabajar cada vez menos. Bajo relaciones humanas libres de explotación, las máquinas permitirían reducir drásticamente el tiempo de trabajo y liberar el tiempo para una vida creativa, emancipada y plena. No es que la tecnología sea neutra, sino que en una sociedad libre de explotación habría inteligencia, conocimientos, libertad y capacidad para construir máquinas que permitiesen a la humanidad liberarse del trabajo.

Entretanto, emerge como tarea la necesidad de afirmar cuán lejos (o cerca) está el trabajo vivo de este trabajo muerto llamado inteligencia artificial.

Heidegger

La pregunta por la técnica es el título de una conferencia que el filósofo Martin Heidegger dictó en la Academia Bávara de Bellas Artes, en el ciclo *Las artes en la época técnica.* Era en 1953 y tenía sesenta y cuatro años. En su juventud había sabido de la publicación de la teoría de la relatividad, que determinaba que las leyes de la física no son las mismas para todos los observadores, y más adelante había visto cómo se completaba la formulación de la física cuántica, que explicaba cómo es el átomo. Y, claro, había visto caer las bombas nucleares.

En su conferencia, Heidegger hace un llamamiento a pensar la esencia de la técnica. Debemos pensar la esencia de la técnica para acceder a ver su verdad. Y esto hay que hacerlo porque la técnica moderna trae peligros. Pensando la esencia de la técnica, accediendo a su verdad, comprenderemos cuáles son esos peligros y podremos hacerles frente.

Y eso ¿cómo se piensa? Pues se piensa dejando de mirar a la tecnología para alcanzar a ver más allá de su uso instrumental. La visión inmediata que tenemos de la técnica es instrumental. Nos relacionamos con las cosas técnicas a partir de lo que se consigue con ellas, poniendo en primer plano su utilidad. Consideramos lo técnico como instrumento, como útil para conseguir determinados fines, como herramienta para el quehacer humano. Un ascensor es un instrumento para elevar personas. Un teléfono es un instrumento para hablar a distancia. Una bomba es un instrumento para matar.

Según Heidegger, todo esto es correcto, pero no es verdadero. Son verdades aparentes, pero no llegan a desvelar la verdad esencial. La pregunta sobre para qué sirve una cosa técnica no nos lleva a ninguna

parte. Y la pregunta sobre su funcionamiento tecnológico tampoco. No sirve preguntarse cómo funciona el motor del ascensor, el hilo de cobre del teléfono o la fisión del núcleo atómico, tal como esas preguntas se formularían en una clase de física.

De hecho, para descubrir la esencia de la técnica hay que separarse, alejarse, de todo esto. «Nunca experimentaremos nuestra relación con la esencia de la técnica mientras nos representemos y dediquemos solo a lo técnico, para apegarnos a ello o para rechazarlo. En todas partes estamos encadenados a la técnica sin que nos podamos librar de ella, tanto si la afirmamos apasionadamente como si la negamos. Sin embargo, cuando del peor modo estamos abandonados a la esencia de la técnica es cuando la consideramos como algo neutral, porque esta representación, a la que hoy se rinde pleitesía de un modo especial, nos hace completamente ciegos para la esencia de la técnica».

Dicho de otra manera, lo evidente está ocultando lo verdadero. La verdad de la técnica, su esencia, no tiene nada que ver con lo técnico. Entonces ¿cuál es la pregunta? La pregunta es cómo la técnica está en el mundo, cómo la técnica explica el mundo. La técnica contiene, lleva implícito, un modo de comprensión del mundo. Ese modo de comprensión del mundo es su verdad verdadera. La esencia de la técnica es su capacidad para dominar la forma en que miramos el mundo y para regular nuestra relación con él. Desocultar esta verdad no es intuitivo. Requiere el esfuerzo de pensar.

Pero el modo de comprensión del mundo que está contenido en la técnica es histórico, va cambiando. Así que Heidegger distingue entre la técnica antigua y la técnica moderna, es decir, entre la visión del mundo subyacente en la una y en la otra. Y pone algunos ejemplos. Antiguamente, el campesino labraba las tierras y eso quería decir cuidar y cultivar. Sembraba las simientes y se abandonaba a las fuerzas del crecimiento, cuidando para que se produjera la germinación. Ahora el campo ha sido absorbido por la industria alimentaria mecanizada. En el antiguo molino de viento las aspas giraban entregadas al soplar del viento, sin interferir con él, sin capturarlo, sin dominarlo. Ahora se interviene el cauce de un río, se domestica el río para construir saltos de agua y generar energía eléctrica. Ya no hay caminos que recorrer en el bosque. Los caminos son ahora recorridos prefijados en un parque temático que dispone de un gran parking a su entrada. Solo hay árboles para talar.

La manera de hacer de la técnica antigua era respetuosa con la naturaleza. Dejaba a la naturaleza brotar, emerger. Ahora la manera de hacer de la técnica moderna es provocadora. Para la técnica moderna, la naturaleza es solo un recurso. El hombre deja de ser su cuidador para actuar de forma provocadora, exigiendo a la naturaleza la entrega de más y más energía: más canteras, más minas, más pozos de petróleo, más árboles talados, más fertilizantes, más pantanos… La técnica moderna provoca a la naturaleza, la excita, la somete y le exige liberar más y más energía que pueda ser acumulada, intercambiada, vendida. Ve la naturaleza como un almacén, como una mera reserva de energía a capturar, almacenar y explotar para el consumo. Esa es la verdad de la técnica moderna.

Pero, con todo, lo peligroso no es este funcionamiento de la técnica: «Lo peligroso no es la técnica […]. La amenaza no le viene al hombre principalmente de que las máquinas y aparatos de la técnica puedan actuar quizás de modo mortífero». Ah, ¿no? ¿Hay algo más amenazante todavía que el actuar mortífero de los aparatos y las máquinas? Para Heidegger sí. Más amenazante todavía que el actuar mortífero es una especie de destino que se ha introducido ya en la esencia del hombre: creyéndose él mismo dominador, dueño y señor del mundo, con el control de la técnica en sus manos, en realidad queda atrapado en esta relación provocadora con la naturaleza y él mismo se convierte en una reserva de energía. Ahora él es un recurso explotado por la técnica, a la que debe suministrar cada vez más y más energía humana. Creía dominar la técnica, disponer de ella a su antojo, y es la técnica la que lo domina a él. La esencia de la técnica moderna es degradante y embrutecedora no solo porque esquilma la naturaleza, sino también porque el hombre, a fuerza de interiorizar una lógica de saqueo, queda él mismo embrutecido y degradado, saqueado, convertido en residuo, en *stock*. Por su voluntad de dominio queda dominado, obligado a no dejar de producir, a producir hasta la extenuación. Se distorsiona la relación que el hombre tiene consigo mismo y se oculta, se ofusca, cualquier otra posibilidad de relación.

El esfuerzo de pensar la técnica pasa por distanciarnos de su uso y preguntarnos cómo nos explica el mundo y qué modo de relación con él nos propone. Heidegger no se preguntaría, por ejemplo, para qué sirven las redes sociales. Es obvio que las redes sociales sirven para relacionarse con otras personas. Y esta afirmación es correcta, pero no desvela la esencia de las redes sociales. No nos lleva a ninguna parte. Lo que hay que preguntarse, para llegar a la verdad, es cómo las

redes sociales ordenan las relaciones sociales. Indagando ahí es como se descubre que las redes sociales son estructuras que proponen un modo de relación social en el que subyace la consideración de que la atención es un recurso a extraer y explotar.

Son estructuras pasivas en tanto que recogen y vehiculan relaciones sociales, pero al mismo tiempo son activas en tanto que las estructuran, las modelan, las conducen. ¿Hacia dónde? Hacia una captura de la atención cada vez más intensiva, porque la atención, en su concepción del mundo, es un recurso a explotar.

Descubrir eso es imprescindible para poder empezar a plantearse mantener una relación libre con las redes sociales. Esa relación libre ya no pasa por subir a las redes contenidos más críticos o alternativos. No pasa por radicalizar las redes sociales. Pasa por liberar mi atención, por restaurar la libertad de mi atención, pero sobre todo pasa por liberarme de concebir la atención (de los otros) como un recurso. Cuanto más me creo reina y señora de las redes, con más seguidores, con más *likes*, con más visibilidad, más estoy siendo explotada por ellas, porque más atención les estoy suministrando. Por eso, por más noble que sea mi causa, no voy a ser yo quien suba un contenido a una red esperando captar la atención de nadie. No voy a degradarme considerando a los demás como un reservorio de atención a explotar. No voy a embrutecerme extenuando mi atención para capturar la atención de otros.

Heidegger estaba muy al caso de las discusiones matemáticas y de las formulaciones tecnológicas de su época. Se carteaba y debatía con científicos. No despreciaba ese conocimiento, pues le ayudaba a pensar. Pero alertaba de que detrás hay algo más esencial. Por eso, la cuestión no es hacer un uso bueno o malo del ascensor, del teléfono o de las redes sociales. La cuestión es liberarse de los modos en los que la técnica nos atrapa, de las lógicas que usa para dominarnos, y para ello hay que zambullirse en el pensar hasta dar con la esencia.

La lógica, el discurso que subyace en la técnica, no cae del cielo. Ha sido generado por humanos. El peligro es que la lógica con la que los humanos han generado la técnica moderna es una lógica bumerán.

Cuando cartografío un territorio para dibujar un mapa estoy fuera del mapa, concibiendo un mapa. Estoy activa, ideando, inventando cómo hacer el mapa, y el mapa es pasivo. Pero, una vez he dibujado el mapa y lo uso, el mapa modela mi concepción del territorio. Veo el territorio a través del mapa, filtrado por el mapa. Ahora el mapa

es activo, es el que tiene la iniciativa. Está ordenando mi visión del territorio, mientras que yo estoy pasiva, asumiendo esa ordenación. Esto es así en todo mapa. Sin embargo, el mapa dibujado en un papel todavía no me conduce, no me lleva por ninguna parte. Pero la cosa cambia con Google Maps, porque Google Maps sí me conduce. ¿Hacia dónde? Hacia mi destino, es obvio. Obvio pero no esencial. Lo esencial es que conduce mi atención hacia la pantalla o hacia las indicaciones sonoras. Captura mi atención, de modo que esta deja de estar en el territorio que está ahí, vivo, fuera del parabrisas. La esencia de Google Maps es cómo ordena mi relación con el territorio o, mejor dicho, cómo la destruye, pues, según su modelo del mundo, el territorio por el que transito es algo completamente ninguneable. No hay nada ahí fuera de la ventanilla. Todo está dentro de la pantalla. La pantalla sabe por dónde voy. Yo no.

Cuando el humano ve el mundo como un contenedor pasivo lleno de cosas cuya energía exprimir, la técnica que desarrollará —coherente con esta visión— será un bumerán que lo convertirá a él mismo en cosa exprimible. Una bomba nuclear sirve para matar personas. Obvio. Pero no es a partir de esta obviedad que podremos liberarnos de su dominio. La esencia de la bomba nuclear es que concibe el átomo como una reserva de energía, como una pila cuya energía está a disposición de quien pueda manipularla, de quien llegue primero. Atrapados en esta concepción, los humanos construyen bombas cuya capacidad puede destruir el planeta entero. Es un sinsentido cuyo eslabón débil no está en la crítica del uso, sino en el desocultamiento de la lógica. Cambiando la lógica, cambiando la relación con el átomo, dejando de considerar el átomo como un *stock* de energía disponible, ya no hay bombas. (Nota filosófica: pero para eso habría que romper con la historia de la metafísica y su concepción del ser entendido como una cosa. Ese es el objetivo último de Heidegger: deconstruir la voluntad de poder y el nihilismo que conlleva la historia de Occidente).

Si Heidegger hubiera conocido la inteligencia artificial, creo que no se habría preguntado para qué sirve, sino a qué relación con el mundo nos conduce. Y creo que la vería como un gigantesco *stock* de conocimiento humano capturado, ordenado y dispuesto a ser vendido. Según la lógica de la inteligencia artificial, el conocimiento humano es un recurso a extraer, someter, privatizar y explotar. Pero la pregunta sobre su lógica deber ser también la indagación sobre su bumerán. Hay que desvelar cómo vuelve el bumerán para no ser golpeados por él.

Cada inteligencia artificial especializa y concreta esta lógica de un modo particular. ChatGPT, por ejemplo, trae implícita la consideración de que la vida no es una página en blanco. Mirar el folio en blanco, la pantalla en blanco, sentir vértigo por ese vacío y sostenerlo hasta que la idea germine es perder el tiempo. La pantalla se debe llenar lo antes posible. No importa que se llene de obviedades, de párrafos estándar, de esquemas manidos. El caso es que se llene instantáneamente, porque no hay tiempo que perder. No importa que te dé siempre la razón, que no se salga ni un ápice de lo más trillado. No importa que el texto se aplane, que se pierda el hilo de las referencias. Que el estilo propio y personal de cada quien esté ausente. Que las ideas ya no vibren con un cuerpo que hoy está así y mañana asá. Que todo quede formateado en párrafos gentiles, redundantes, sin ruido, sin pálpito, sin intensidad… Nada de eso importa, porque ChatGPT es un congelador de conocimiento, un dispositivo para rellenar pantallas que tengo disponible siempre ahí y que me ahorra tiempo, igual que las lentejas congeladas que saco al medio día para no cocinar.

Pero esto tiene un bumerán. ¿Cómo ordena ChatGPT mi relación con el conocimiento, con ese conocimiento que puedo explotar, pero que ya no tengo que cuidar ni atender, ni cultivar? ChatGPT es pasivo, porque recibe mis preguntas. Pero es activo, porque me atrapa en la ficción de una conversación. Le pregunté, me contestó. Mucho mejor que conversar con un humano de carne y hueso, que se va por las ramas, que llega con sus propias neurosis. Mucho mejor que leer un conjunto de textos humanos fragmentados, parciales, contradictorios, desestructurados entre sí, que tengo que contextualizar, componer, comprender. Mucho mejor que enfrentarme al vacío de pensar qué es lo que quiero pensar. Me creo que ChatGPT está en mis manos porque contesta mis preguntas, pero en realidad soy yo quien bailo al son de su lógica: la indeterminación de una página en blanco no aporta nada y no tiene sentido detenerse ante ella, no vaya a ser que perdamos el tiempo.

Entonces ¿Heidegger diría que tenemos que alejarnos lo más posible del ascensor, del teléfono, de las redes sociales, de ChatGPT y, por supuesto, de la bomba nuclear? Para nada. No debemos condenar la técnica, como si fuera obra del diablo. Todo lo contrario: «Cuando nos abrimos a la esencia de la técnica nos encontramos tomados inesperadamente por un reclamo libertador». Y cita los conocidos versos de Hölderlin: «Donde hay peligro crece también lo salvador».

La esencia de la técnica es ambigua: alberga el peligro y también lo liberador. Lo salvador, lo liberador, no se puede captar inmediatamente sin preparación, pero, si miramos atentamente el peligro, veremos también cómo brota, cómo surge lo que salva. Con eso no estamos ya salvadas, porque hay que cuidar el crecimiento de eso que salva. Pero empezamos a ver el brillo de su luz. Ahora bien, para vislumbrar ese brillo es preciso no quitar ojo del peligro. Hay que enfocar siempre al peligro.

Enfocar al peligro es pensar. Es habitar en medio de la naturaleza sin someterla ni dominarla. Es tener una relación más libre, menos instrumental, con la técnica. Es reconocer que no somos dueños y señores de nada. Es no permanecer embelesados con lo técnico. Es separarse, tomar distancia y pararse a meditar.

En sus propias palabras: «Podemos decir sí al inevitable uso de los objetos técnicos y podemos a la vez decirles no en la medida en que rehusamos que nos requieran de modo tan exclusivo, que dobleguen, confundan y, finalmente, devasten nuestra esencia. Quisiera denominar esta actitud que dice simultáneamente "sí" y "no" al mundo técnico con una antigua palabra: la Serenidad para con las cosas».

En las mías: dejar que Google Maps, las redes sociales y ChatGPT descansen.

Simondon

Gilbert Simondon defendió su tesis el 19 de abril de 1958. La tesis consistía en dos trabajos: *La individuación a la luz de las nociones de forma y de información* y *Sobre el modo de existencia de los objetos técnicos.* Era un filósofo profundamente crítico con la separación que la cultura ha ido forjando entre las ciencias y las letras. Además de en filosofía, se había formado en psicología y en tecnologías. Seis meses antes de leer su tesis, la Unión Soviética había lanzado el primer satélite artificial que orbitó alrededor de la Tierra, el Sputnik I. Después de eso ya no se podía seguir pensando la técnica con las ideas del siglo XIX. Ya no podía ser reducida a mero instrumento, a un medio para conseguir un fin, como el arado, que es un instrumento que sirve para arar, o la batidora eléctrica, que es un instrumento que sirve para triturar. El Sputnik I, la punta de lanza de la técnica en 1957, no era un instrumento. No servía para nada. No se podía usar. Y, sin embargo, no era inútil.

La intención de Simondon al aplicar una mirada a la vez filosófica e ingenieril sobre la técnica era suscitar una toma de conciencia sobre los objetos técnicos. «La cultura se ha constituido en sistema de defensa contra las técnicas; ahora bien, esta defensa se presenta como una defensa del hombre, suponiendo que los objetos técnicos no contienen realidad humana».

La cultura se defiende de lo técnico, lo expulsa de su ámbito porque lo considera amenazante. Imagina dentro de la técnica un alma malévola que alberga malas intenciones contra el humano. Hay que mantener la técnica a distancia, lo más lejos posible, porque deshumaniza. Podemos parodiar una escena imaginaria (pero no tanto) en

la que, después de una cena en un restaurante, a la hora de repartir la cuenta alguien dice: «Hazlo tú, que a mí se me dan muy mal las mates. Bueno, no es que se me den mal. Es que las odio. Ya sabes, soy de letras». Si Simondon estuviera ahí, se le abrirían las carnes. ¿Cómo es posible que, en una sociedad tan tecnificada como la nuestra y siendo la técnica inventada y producida por humanos, alguien se ufane de su ignorancia? Esto no ocurre con la cultura de letras. Nadie presume de escribir con faltas de ortografía. La cultura ignora sistemáticamente que la técnica es una realidad humana. Se puede ser una persona culta sin tener ni un ápice de cultura técnica.

Para Simondon, «la oposición que se ha erigido entre la cultura y la técnica, entre el hombre y la máquina, es falsa y sin fundamentos. Solo recubre ignorancia o resentimiento. Enmascara detrás de un humanismo fácil una realidad rica en esfuerzos humanos y en fuerzas naturales, y que constituye el mundo de los objetos técnicos, mediadores entre la naturaleza y el hombre». Ese humanismo fácil promociona el desconocimiento de la máquina. Pero eso no es culpa de la máquina. El humanismo fácil desprecia lo que en la máquina hay de humano. La máquina es vista como un otro ajeno, desdeñable e indigno, igual que los esclavos en la época de la esclavitud. Ante este desequilibrio, la filosofía tiene la obligación de cumplir un deber análogo al que cumplió en la abolición de la esclavitud cuando afirmó el valor de toda persona humana. El filósofo-tecnólogo tiene que erigirse en representante de los seres técnicos y defenderlos contra ese humanismo fácil que los desprecia y los humilla. Tiene que hablar en su nombre, alzar su voz entre aquellos que elaboran la cultura para devolver a los seres técnicos su dignidad y su significación. Tiene que reconciliar la cultura con la técnica.

Cuando Simondon habla de que hay que abolir la esclavitud de las máquinas no está usando una imagen metafórica, sino que se refiere a una realidad directa. Si entramos en una vivienda cualquiera, encontraremos objetos estéticos, como el cuadro colgado en la pared, el jarrón en la mesita de la esquina… Encontraremos objetos sagrados, como las velas junto a las fotografías, los recuerdos de los que ya no están… Pero los objetos técnicos están escondidos, disimulados. Los cables y las cañerías, empotrados en las paredes; los enchufes junto a los zócalos, que no se vean mucho; los electrodomésticos panelados, que parezcan muebles… Al igual que en las casas de los ricos hay una entrada para el servicio separada de la de los señores, al igual que en las viejas mansiones las cocinas están en los sótanos —porque,

pudiendo ver salones y pianos, ¿quién querría ver sacos de patatas y fogones?—, al igual que la servidumbre, los criados, son inferiores e indignos, de la misma manera el hombre trata a la máquina como a una esclava: siempre disponible, siempre a su servicio y sin poseer ella misma ni significación ni valores.

Pero esta relación de dominación de lo humano sobre lo técnico no solo aliena a las máquinas en tanto que las despoja de su verdadero ser, sino que aliena también al humano en tanto que desprecia su propia humanidad, lo que de humano hay dentro de la máquina. Es como si el humano se despreciara a sí mismo, despreciara su huella cristalizada en la máquina, despreciara su propia inventiva. Por ignorancia, el humano se siente amenazado por la máquina y busca seguridad dominándola, sometiéndola. Y así se convierte en dominador. Pero debe liberarse de esa relación de dominación sobre las máquinas y sobre la naturaleza. El humano no está por encima. No hay nada por encima de nada. Todo (naturaleza, humanos y máquinas) está en el mismo plano. El humano debe tomar conciencia de que no es ni superior ni inferior: vive entre máquinas, en horizontalidad con ellas. Y les debe respeto.

Mientras preparaba su tesis doctoral, Simondon daba clases de filosofía en un liceo y se preguntaba qué sería una formación humana completa. Se preguntaba cómo debería ser una educación que despertara la imaginación de los chicos y chicas y las preparara para inventar cómo resolver ese problema siempre nuevo que es la vida. Observaba que el liceo respondía a las necesidades de educación de la burguesía. La burguesía no va a hacer trabajo manual, porque mancharse las manos es indigno y propio de sirvientes u obreros. La burguesía va a dar órdenes (prescribir, registrar, legislar…), así que recibe una enseñanza altamente simbólica, especialmente en lo referido al lenguaje. Por su parte, el obrero recibe una educación de oficio y su talento se manifestará en su capacidad para repetir un gesto eficaz. Esta educación, que reproduce la jerarquía de las clases sociales, debe ser sustituida por una educación unitaria, integradora, igual para todo el mundo por lo menos hasta los dieciséis años.

Defendía una educación que transformase al albañil en constructor, al obrero en ingeniero y al burgués en un trabajador que manipule la materia con la misma nobleza con la que practicaría un deporte. Y se arremangó para impartirla, realizando con su alumnado talleres en los que armaban y desarmaban motores de automóvil o construían aparatos de televisión. Tras la experiencia, expuso las dificultades.

Treinta alumnos por grupo son demasiados. Hace falta un espacio adecuado: tener que desmontar un motor en un aula llena de sillas y pupitres es penoso. Se necesitan presupuesto, recursos y apoyo por parte de la dirección del centro; él mismo se pateaba los desguaces buscando los componentes que necesitaba. Y hace falta un cambio de mentalidad para superar los prejuicios y las críticas a las que tuvo que hacer frente: que si expones al alumnado a tareas demasiado peligrosas, que si los enfrentas a una complejidad que a su edad no pueden comprender…

Simondon reivindica el taller. No desdeña el laboratorio científico, pero el lugar donde vive la técnica es el taller. Ahí es donde crece la inventiva. Un profesor de filosofía dando talleres de tecnología con una propuesta clara: es preciso formar un equipo que esté compuesto por la profesora de filosofía, la profesora de historia y la profesora de ciencias, además de la tecnóloga. Se necesitan estas cuatro miradas para comprender y transmitir al alumnado la profunda dimensión de lo técnico.

Aunque nuestras escuelas actuales proporcionan la misma enseñanza primaria y secundaria a todas las personas, al menos sobre el papel, si Simondon entrara en un instituto, quedaría profundamente decepcionado: las estudiantes pasan horas y horas sumidas en la sequedad de los ejercicios y los escasos momentos de manipulación de la materia se presentan como pasatiempo en forma de manualidades. Nada de pensamiento aplicado. La tecnología profunda, la que expresa el sentido del trabajo de la cultura popular, el sentido del saber de la cultura burguesa y el sentido del acto heroico de la cultura nobiliaria, la que compone y unifica estas tres culturas, está barrida del sistema educativo.

Hay que salvar al objeto técnico, pero ¿qué es lo que hay que salvar en él? La respuesta de Simondon es contundente: hay que salvar su tecnicidad. Salvar la tecnicidad de la técnica parece un juego de palabras que no dice nada, una perogrullada. Pero no lo es. La tecnicidad no es lo mismo que la técnica. La tecnicidad es el saber hacer concreto, eficaz y materializado. En el origen, el único ser portador de tecnicidad era el humano. Pero ahora también hay máquinas que tienen tecnicidad, que saben hacer algo concreto de manera eficaz. El meollo de la cuestión es que cuando el humano no comprende la tecnicidad que existe en cada máquina, cuando la ignora o la desprecia, entonces la técnica se expresa solo como dominación, como dominación del humano sobre la máquina, cuando el objeto técnico se usa

instrumentalmente y ocupa el lugar del esclavo. Como este, deberá obedecer sin fallas, ser fiel, no mostrar espontaneidad, no entrar en rebelión, no manifestar su vida interior, su mecanismo, sus dificultades. O como dominación de la máquina sobre lo humano, cuando el humano, en sus delirios tecnocráticos, idolatra la tecnología como a un tótem de poder.

La técnica surge cuando en un determinado umbral antropológico el humano se separa del mundo. Hasta ese momento todo era todo y todo estaba vinculado con todo. Había unidad. Pero en un determinado momento la unidad entre el mundo y el humano se resquebraja. El humano rompe con el mundo y se separa de la naturaleza. Como si un personaje se saliera del cuadro en el que está pintado, la figura, el humano, se separa del fondo, la naturaleza. A partir de ahí la técnica va a estar en medio, va a ser la mediadora entre figura y fondo. La tecnicidad, que está en medio, tiene parte de naturaleza y parte de humanidad. Es un modo de relación entre humanos y naturaleza, mediada por la técnica. El otro modo de relación será el religioso. Y a medio camino entre ambos modos estará el arte.

La tecnicidad es espacio común, espacio de encuentro entre el humano y la máquina. Si las preposiciones marcan la pre-posición, la posición relativa de unas cosas respecto a otras, en el pensamiento antiautoritario de Simondon la preposición estrella es el «entre». Máquinas entre humanos, humanos entre máquinas. No a las relaciones verticales. Sí a la horizontalidad. Naturaleza, máquinas y humanos, todo en el mismo plano.

Entonces ¿hay que salvar a todas las máquinas? No exactamente. Hay que salvar más a las que tienen más tecnicidad, porque son las más congruentes con lo humano, con la tecnicidad humana. ¿Y eso cómo se sabe? Simondon apunta tres criterios: apertura, indeterminación y lenguaje.

El objeto técnico no es una recompensa para el hombre. No es como un caramelo para el niño que se ha portado bien, sin que haya vínculo ninguno entre el caramelo y el comportamiento. No es como el premio que recibe el muchacho que ha disparado con puntería en la caseta de feria. No es un trofeo. Es la traducción, el registro fiel del trabajo humano. Es la cristalización de una larga serie de esfuerzos dirigidos por una intención sostenida y reflexionados por una voluntad inteligente. Si se sigue la línea de ese registro fiel, la mayor tecnicidad, la mayor perfección, se manifiesta en la apertura, la indeterminación y el lenguaje. Porque estas tres cualidades resuenan con lo humano.

Apertura es mostrar el interior. Quitarse la máscara. Despojarse de todos los maquillajes que esconden el funcionamiento. El objeto técnico no es bello porque esté adornado con elementos que nada tienen que ver con su ser, sino porque se expone, se manifiesta con autenticidad. Se expresa tal como es. Una locomotora antigua es más auténtica que una moderna, porque todo su funcionamiento está al descubierto. Todo lo contrario a un tren de alta velocidad, revestido por una carcasa reluciente que opaca su interior y que parece decirnos: «Monta y úsame, pero mejor si no ves nada». Un camión tiene más belleza técnica y más autenticidad que un coche que disimula su técnica debajo de mil embellecedores. Una moto, con su motor a la vista, o una bicicleta, con sus cadenas y sus engranajes llenos de grasa, son auténticas.

Pero apertura también es una disposición interior comprensible en la que los componentes se entienden y pueden ser sustituidos por otros nuevos o superiores en caso de avería o de mejora. Si cuando falla un componente he de tirar el aparato entero, entonces no hay apertura. A un ordenador de sobremesa se le puede sustituir la pantalla por otra más grande o mejor. A un portátil se le puede añadir una segunda pantalla. Son máquinas abiertas. Con una *tablet* no se puede hacer eso. Está cerrada. Sin apertura no puede haber evolución. La tecnicidad está capada. Es una tecnicidad frustrada.

Indeterminación significa que en la máquina hay algo aún por definir. Es apertura al medio. Es sensibilidad hacia lo que ocurre en el exterior y capacidad de respuesta. Es todo lo contrario del automatismo cerrado. Para Simondon un robot autómata es de una tecnicidad muy baja. No hay mucho que salvar en él. El robot autómata que va por libre no está diseñado para el entre. No necesita de nada ni de nadie. No necesita de otros seres técnicos, ni maquínicos ni humanos. La indeterminación de la máquina es lo que permite al humano interpretarla y relacionarse con ella. Una máquina completamente definida, cerrada en sí misma, impide la relación. El automatismo no hace a la máquina más perfecta. Es al revés. La indeterminación abre un espacio de incertidumbre que requiere de la relación y de la interpretación. Da margen de maniobra. Eleva la tecnicidad.

Y, para que pueda haber relación, humano y máquina deben compartir un lenguaje común. Debe haber un simbolismo común que permita la comunicación, que permita que ambos mantengan una especie de relación social. Debe haber una sinergia, una interfaz congruente entre ambos. La máquina debe poder darse a conocer y el

humano debe conocerla para poder guiarla. Tiene que haber un canal de comunicación de ida y vuelta. «Así, el objeto técnico es inventado, pensado y querido, asumido por un sujeto humano».

Apertura, indeterminación y lenguaje son los criterios de la alta tecnicidad porque son las condiciones para que el operador humano tenga cabida. Máquinas que nos dan y nos exigen servicios, como una amiga. Máquinas que funcionan en ese entre, en ese único modo de relación libre para todas las partes. Máquinas siempre en curso, nunca dadas de una vez por todas.

Aun cuando hay voluntad, esa toma de conciencia que quiere suscitar Simondon es difícil, porque raspar la tecnicidad cuesta. La relación que él tuvo con la técnica es muy distinta de la que tenemos nosotras. Si nuestra experiencia no resuena con su amor a las máquinas, es porque la nuestra es una experiencia condicionada por los intereses mercantiles en una sociedad de consumo. Cuando las máquinas están diseñadas para ser vendidas como objeto de consumo, empiezan a cerrarse y a estetizarse. Uno de los emblemas de este proceso es Apple. Sus ordenadores, con una integración blindada de arriba abajo, con su propio sistema operativo, sus propios programas y su propio hardware, son joyas de la informática. Objetos de prestigio estetizados con una gama de colores espectacular, con carcasas de aluminio futuristas. Bonitos diseños y cero apertura. ¿Bonitos? Para Simondon no, porque se trata de un bonito postizo, un bonito de pega que va en detrimento de la auténtica belleza técnica.

Cuando las empresas se hacen con la técnica, cuando se imponen los criterios de mercado, avanza la técnica pero retrocede la tecnicidad. No te plantees reparar una Roomba de esas que barren el suelo mientras estás fuera de casa. En el supuesto de que puedas abrirla, que ya es mucho suponer, no encontrarás nada inteligible en su interior. La Roomba no está diseñada para que la entiendas, la quieras y la asumas. Está diseñada para que la estrujes y cuando ya no sirva la tires. Amo y esclavo. Y así es nuestra experiencia tecnológica, forjada por encuentros indiferentes con objetos técnicos carentes de tecnicidad que no tienen nada que decirnos, salvo: «Úsame, soy tu sirviente». ¿Cómo vamos a amar máquinas así?

Simondon no llegó a ver el auge de la inteligencia artificial. Vivió en la época en la que las máquinas de energía, como las calderas o los motores, empezaban a dar paso a las máquinas de información, como la televisión o el radar. No llegó a ver el enorme despliegue de

las máquinas virtuales, hechas de puro código soportado por microchips. Sin embargo, las preguntas y los criterios que desarrolló para comprender lo técnico siguen vibrando. Y la toma de conciencia que quiso motivar sigue siendo más necesaria que nunca.

Muy interesado por la cibernética, la definió como la ciencia de las relaciones. No la consideró una amenaza para el humano, sino una oportunidad para hacer comunidades de intercambios con las máquinas, «un acoplamiento del hombre y de la máquina en la misma unidad funcional». Humanos tecnificados y máquinas humanizadas, y ambos siempre mirando al mundo. El humano como mediador en una sociedad habitada por máquinas. Y pone el ejemplo del director de orquesta: «Lejos de ser el vigilante de una tropa de esclavos, el hombre es el organizador permanente de una sociedad de objetos técnicos que tienen necesidad de él como los músicos tienen necesidad del director de orquesta». El director puede dirigir la orquesta de músicos porque es capaz de tocar como cada uno de ellos. Pero no manda, solo organiza. Para cada intérprete, el director es el medio para captar al grupo. Es el indicativo del dinamismo del grupo, su forma en movimiento. Cada intérprete percibe y se inscribe en el grupo a través del director. De igual modo, dice Simondon, el hombre tiene como función ser el coordinador e inventor permanente de las máquinas que están alrededor de él.

No hay nada malo en una humanidad que dirige una orquesta de inteligencias artificiales que interpretan el son del mundo con una tecnicidad de la buena. La cuestión es si no será la inteligencia artificial la que está dirigiendo la orquesta de humanos.

Actitud

Carta a Amador

Querido Amador, ¿cómo estáis?

Ya me he leído el libro de *La actitud hacker*, el de Carlo Milani. No veas lo que me ha gustado. Nada que ver con el de la ética hacker, el de Himanen, y otros similares, que hablan desde una cultura meritocrática anglosajona que no es la nuestra. Este de Milani es una reflexión sobre lo hacker desde la manera de vivir mediterránea, más como nuestros hacklabs, nuestros hackmeetings… Parecía que todo lo hacker ya estaba escrito, pero no. Todavía se pueden decir más cosas ;-)

Y la idea clave es muy buena. Lo hacker es una actitud. Teníamos la ética y teníamos la comunidad. Y ahora tenemos también la actitud. Y va explicando en qué consiste. Lástima que el título pueda tirar atrás a muchas lectoras, porque es un libro que podría ser de lectura masiva. De hecho, está dirigido a gente no especializada. En fin, que muchas gracias.

He entendido mejor por qué me gusta tanto dar talleres de robótica en los coles. Me gusta porque en un aula de primaria el robot es funcionamiento sin uso. Lo ponemos a funcionar, pero no lo usamos. No es ni una herramienta ni un juguete. De mayores usarán las herramientas para el trabajo y los juguetes para el ocio. Pero el robot está ahí en medio, entre una cosa y la otra sin ser ninguna de las dos, como un amigo que nos ayuda a comprender nuestra humanidad.

Estuve en un encuentro en Barcelona para poner en común perspectivas sobre la inteligencia artificial. Había hackers, militantes por los derechos humanos, las de la cultura libre, gente que está en políticas públicas… No te puedo pasar la web, porque no la hay. Ja, ja, ja. Me perdí el taller de instalación y reentrenamiento de una IA generativa en el ordenador portátil. Argggggggggg, no pude ir.

Algunos decían que la IA va a ser como el nuevo sistema operativo. Igual que ahora tenemos el Ubuntu o el Windows y ahí encima lo hacemos todo, pues de fondo vamos a interactuar con una IA, que va a ser la encargada de llamar a otras IA más específicas dependiendo de si quiero hacer un cálculo, una búsqueda, una traducción, escribir un texto, jugar…

Sé que no hay consensos hackers sobre esto de la IA. Hay gente que está más por jugar y otra gente está más por romper la baraja. Pero el tema se está moviendo. Saldrán cosas. Tienen que salir :-)

Bueno, a ver si la próxima es un paseo por el Retiro.

Abrazos.

7. Hacer

Hackers

En los años sesenta, los años de la guerra de Vietnam, una tecnoélite contracultural empezaba a trastear con una nueva disciplina académica. Eran hackers, jóvenes que programaban ordenadores con el deseo de aprender y de mejorar las tecnologías existentes.

Aunque con toda la razón se suele decir que el gran invento hacker es Internet, su legado tiene otras dos patas no menos importantes: su modo de hacer las cosas, la ética hacker, y su modo de cooperar, la comunidad hacker.

Cuando el *hacking* llegó a nuestras tierras, en los últimos años del siglo XX, los movimientos sociales, simplificando mucho, se polarizaron en dos visiones opuestas. Por una parte estaban los que rechazaban el uso de las nuevas tecnologías por considerarlas un instrumento capitalista para perpetuar las relaciones de poder entre las clases sociales y las relaciones imperialistas entre los Estados del Norte y los del Sur global. Por otra parte estaban los que defendían que los movimientos sociales debían apropiarse de esas nuevas tecnologías y aliarse con ellas.

¿Y eso de aliarse cómo se haría? Otra vez simplificando, había dos visiones.

Una visión respondía a la consigna «dar voz a los sin voz» y proponía construir medios de comunicación telemáticos alternativos. Básicamente, utilizar Internet como una herramienta para el desarrollo de las comunicaciones, para hacerlas más rápidas, más baratas y más eficientes. Es decir, usar Internet como un altavoz, aprovechando su capacidad para amplificar los mensajes alternativos, las críticas y

las denuncias. El primer hito de esta línea de acción tuvo lugar en 1993, cuando la campaña 50 Años Bastan, organizada por el movimiento antiglobalización contra las instituciones de Bretton Woods (Fondo Monetario Internacional y Banco Mundial), decidió montar una infraestructura telemática para tener un instrumento de comunicación propio.

Otra visión respondía a la consigna «construyamos Internet» y proponía abrazar los retos tecnosociales que planteaban las nuevas tecnologías y meterse hasta el fondo de la red para experimentar, para llevar al límite las posibilidades de construir espacios de autonomía. Básicamente, tomar la delantera en la construcción de Internet para poblarla de zonas liberadas, de espacios sociales libres de las estructuras de control y de poder. Producir formas de vida y de socialidad —fuera de las redes, pero también dentro— refractarias al mercado y al mando. El primer hito de esta línea de acción fue la fundación, en 1999, de sindominio.net, un servidor de Internet autónomo y autogestionado por una comunidad horizontal organizada en asamblea permanente.

A primera vista, puede parecer que esta segunda visión fracasó estrepitosamente, aunque eso, aparte de que requeriría una evaluación más pormenorizada, no es ningún argumento en su contra. No se puede comprender un movimiento social mirando solo desde fuera, atendiendo a lo que consigue. Para ver las posibilidades que abre hay que mirar desde dentro. Las luchas por construir Internet también han conseguido cosas. Aunque ahora cueste verlo, han levantado muchos obstáculos para impedir que los poderosos hicieran de las suyas. Han saboteado planes corporativos y han condicionado la hoja de ruta de grandes corporaciones para que la Internet abierta, libre y neutral no sea una quimera, sino una realidad tangible (por supuesto, no ideal) por cuya defensa hay que seguir luchando. (Animo a la lectora a no ofuscarse identificando Internet con las redes sociales comerciales. La red de redes es mucho más que unos cuantos servicios privativos).

En su día, «construyamos Internet» criticaba a «dar voz a los sin voz». Les decía: «Estáis usando Internet, pero no estáis haciendo Internet». Las nuevas tecnologías permiten nuevas prácticas, pero el simple contacto con las nuevas tecnologías no es suficiente para modificar prácticas profundamente arraigadas en las viejas formas de hacer política. No se trata de subir a Internet contenidos alternativos o críticos. Los entornos digitales van a ceñir o expandir la vida tanto

como lo hace el urbanismo. Así que no vale decir que Internet es solo una herramienta para la comunicación, igual que no vale decir que el urbanismo es solo una herramienta para el uso ordenado del espacio. Que podamos operar como máquinas-herramienta encarnadas que expresan, comunican y producen junto y en conexión con otras, en red, plantea nuevas ambigüedades, otras formas de placer y también de poder, otros lenguajes, otros sufrimientos. Hay que experimentar hasta el fondo las posibilidades de construir sociedades *online* más libres que las que están hechas de dinero, de poder, de cemento y acero. Hay que implicarse hasta las trancas. Hay que dislocar hasta el límite. Y eso no se puede hacer si priman las razones instrumentales.

Para «construyamos Internet» había un línea roja: el software ha de ser libre. Solo hay un software, y es el software libre. Lo que no se pueda hacer con software libre, sencillamente, no se hace. En el supuesto de que se quisiera hacer, habría que ponerse manos a la obra para construir ese software que se necesita y no se tiene. Y eso se puede hacer. Lo podemos hacer si colaboramos, porque el conocimiento está disponible.

Para la ética comunitaria hacker hay principios irrenunciables: el software libre se construye en comunidad y la comunidad es la garante de la gobernanza colectiva de ese software, que es un bien común. El conocimiento debe ser compartido y los desarrollos individuales deben ser devueltos al común. La inteligencia es colectiva y privatizar el conocimiento es matar la comunidad. La comunidad es imprescindible para que la inteligencia social circule y se retroalimente, según una lógica de cooperación. (Principios que resuenan con el intelecto general de Marx, las lógicas no extractivas de Heidegger o las relaciones horizontales y no instrumentales de Simondon).

Así que, en resumen, con la llegada de las nuevas tecnologías a finales del siglo pasado, los movimientos sociales desplegaron dos lógicas. Una consistía en usar las tecnologías disponibles, libres o privativas, y hacer un uso convencional o disruptivo de las mismas (por ejemplo, usar redes sociales comerciales para autoorganizar y extender las luchas), alterando los usos previstos de esas herramientas. La otra consistía en extender las tecnologías libres (hardware y software) expandiéndolas bajo una lógica comunal, de bien común, de propiedad y gobernanza colectiva. La transformación fundamental radicaba en romper las lógicas de propiedad privada sobre las tecnologías. Y para ello, el acceso al código y al conocimiento que albergan es imprescindible e irrenunciable.

Cabe decir que estas dos visiones tenían puntos de encuentro. «Construyamos Internet» no se oponía a «dar voz a los sin voz». Es solo que le parecía que las condiciones permitían llevar la experimentación más allá. El proyecto Indymedia, por ejemplo, fue un cruce de caminos. Vinculado al movimiento antiglobalización, creó una red de nodos *online* que informaban sobre temas políticos y sociales, utilizando un proceso de publicación abierto en el que cualquier persona podía publicar. Para ello se programó un software libre que utilizaba unas innovaciones tecnológicas que se habían desarrollado para el lenguaje de programación PHP. Pero, aunque hubo confluencias, las dos visiones palpitaban con corazones distintos.

Ahora, ante la inteligencia artificial, más allá del movimiento hacker, la gente también tiene posicionamientos distintos. Uno es que todo esto es un bluf, un globo que se pinchará por sí solo cuando se evidencie que las expectativas no responden a la realidad de lo que realmente puede ofrecer este conjunto de tecnologías. Bastaría con no dejarse arrastrar por las expectativas y esperar a que el globo se deshinche. Otro es que la inteligencia artificial es una amenaza muy seria contra la justicia y la equidad, que abrirá escenarios de exclusión social muy agresivos y dinámicas totalitarias. Habría que oponerse frontalmente a ella, resistirse y luchar por su abolición. Un tercero es que las máquinas inteligentes son una amenaza para la humanidad en su conjunto y que estamos ante un riesgo existencial muy elevado, porque tendrán supremacía sobre lo humano y capacidad para tomar el control de la vida. Sería necesario establecer una regulación, aunque es complicado hacerlo, porque esta tecnología se encuentra en sus inicios.

Este libro es un intento de ralentizar, por el momento, la adhesión a una visión, sea cual sea. Es un intento de detener, provisionalmente, la opinión y darnos más elementos para poder construir un pensamiento propio. No adherirnos a pensamientos ya hechos, ya armados, hasta haber abierto los espacios para montar uno propio. Desembotar la imaginación. No decantarse a la primera de cambio. Y no porque esas visiones sean malas, sino porque, además de esperar, resistirse o temerla, tiene que haber más maneras de relacionarse con ella. Tiene que haber maneras propias, activas, creativas, desafiantes, apasionadas de a la vez resistirse y generar conflictos, y al mismo tiempo construir y abrir posibilidades. Crear y a la vez destruir. Tiene que haber una manera más hacker de relacionarse con todo esto. Seguro que tiene que haberla.

Obviamente, no puedo exponer cuál podría ser la propuesta hacker, más que nada porque el movimiento es plural, porque no llego a todos sus rincones y porque no hay una propuesta unificada (ni falta que hace). Lo que sí puedo es esbozar los horizontes de deseos y las líneas de conflicto que se comparten en encuentros y en espacios de acción local con los que tengo relación.

Lo primero sería dejar de considerar la inteligencia artificial como un monolito. Utilizar lo menos posible el término «inteligencia artificial» y poder nombrar cada cosa según lo que creemos que es. Entender para desmitificar. Si la imagen mental que tenemos es la de un gran monolito envuelto todo él por un enorme celofán, entonces la crítica se hace sencilla, se hace simple. Se critica el bloque entero y ya está. Es simple, pero no funciona. Con ello se contribuye a hacer el bloque aún más grande, al tiempo que se hace imposible ver o imaginar las disidencias, las rupturas, las apropiaciones… Hay que formar un bloque muy grande para poder enfrentarse a otro bloque grande, y la actual relación de fuerzas no es favorable. Totalizar es problemático.

Como en el juego infantil de las palabras prohibidas en el que no se puede decir ni sí ni no, ni blanco ni negro, ni oro ni plata…, vamos a ver si podemos hablar de inteligencia artificial sin usar el término. Vamos a ver si podemos elaborar un pensamiento propio que rasgue el celofán, que trocee el bloque, amplíe el repertorio de ideas con las que pensar y sitúe cada ensamblaje humano-máquina en su particularidad, atendiendo a su tecnología, a su ideología y a cómo se acopla en un entramado social concreto. Decir que AlphaZero, Apertium, Google Maps, Viogén, Alexa y Mini son inteligencias artificiales es como no decir nada. Es más, es como hacer el caldo gordo a la ideología. Habría que poder decir algo más concreto, más propio, más fino, de cada una de ellas. Romper el bloque. Separarlas. No reducir la complejidad, sino aumentarla.

Para aumentar la complejidad, para hacer estallar las posibilidades, hace falta *hacking*. Hacen falta espacios de experimentación. Hay que activar dinámicas de taller, porque el trasteo y la compartición de conocimientos siguen siendo necesarios. No se puede comprender bien lo que no se conoce. Para ver por dentro, hay que meterse dentro. Hay que trastear con las tecnologías. Hay que equivocarse y aprender del error. Hay que generar conocimiento propio.

Flota la idea de que hacen falta muchos recursos de hardware para entrenar redes neuronales profundas. Y eso es cierto, pero matizable. Los modelos megalómanos tipo ChatGPT, que aspiran a resolverlo todo y que llevan en sus entrañas la matriz de un monopolio totalizador, no son los únicos posibles ni deseables.

A día de hoy es viable y relativamente sencillo, por ejemplo, instalar un chatbot en un ordenador personal de gama media. GPT4All es un software que puedes instalar en tu ordenador para tener funcionando un modelo de lenguaje en tu portátil. Proporciona privacidad total, puesto que los datos no salen de tu ordenador. Funciona sin conexión a Internet, así que no necesitas servidores en centros de datos funcionando las veinticuatro horas del día. Es software libre y su código está en GitHub. Se desarrolla según un modelo de comunidad. Y permite usar diversos modelos de lenguaje, con diferentes tamaños, especializaciones, requisitos y licencias.

Aunque una vez entrenado un modelo es posible ponerlo a funcionar en un ordenador personal, para entrenarlo hacen falta muchos recursos hardware. Cierto. Pero los recursos existen. El supercomputador Marenostrum, en el Barcelona Supercomputing Center, es una infraestructura pública del Estado español que está disponible para proyectos científicos europeos, y está en el top de los tres mejores ordenadores europeos. El gobierno ha aprobado un plan para reforzar su capacidad de procesamiento orientada a inteligencia artificial y desde el proyecto ALIA se van a generar modelos de lenguaje en todas las lenguas cooficiales del Estado. Dicen que los modelos serán públicos, abiertos y transparentes. Habrá que ver cómo se concreta esa voluntad.

Aquí aplica la vieja consigna de que lo que se financia con dinero público sí o sí tiene que ser software libre. No hay nada, *a priori,* que nos impida disponer de modelos de lenguaje públicos funcionando con software libre en nuestros ordenadores personales. Aunque, a lo mejor, hay que organizar algo de conflicto para que lo público siga siendo público y hay que reactivar el espíritu hacker investigador y universitario para armar alianzas público-comunitarias bajo el modelo de los bienes comunes. A la vez hacer alianzas, a la vez generar conflicto y a la vez reinventar lo libre.

La discusión sobre qué significa software libre (que no gratis) sigue siendo muy relevante. Decir que algo es software libre es darle un significado muy preciso que ha sido meticulosamente definido y

defendido a lo largo de décadas por las comunidades hacker y que se concreta en los cuatro derechos sobre el código: derecho a usar, a estudiar, a mejorar y a compartir. Pero la presión de lo privativo se disfraza con muchas caretas. Por ejemplo, la empresa Meta, propiedad de Mark Zuckerberg, ha creado Llama. Llama es un modelo de lenguaje con el que Meta quiere competir con ChatGPT. Y lo quiere hacer pasar por software libre, cuando en realidad no lo es. Proporciona acceso a algunas partes del código, pero está lleno de restricciones.

Más allá de los tecnicismos sobre licencias de software, Meta hace eso porque «software libre» es un concepto que mola. Tras años y años de luchas por el derecho a usar, estudiar, mejorar y compartir el código, ahora «software libre» es un término de prestigio que las corporaciones usan para obtener el respaldo de las comunidades de programadoras. Pero hay que estar al quite. Las empresas no pueden etiquetar un producto como software libre cuando no lo es, porque eso es engañar a las usuarias y es fragmentar las comunidades. La defensa del software libre sigue siendo necesaria. Y a ello hay que añadir, ahora, la necesidad de reinventar qué va a ser lo libre respecto a los datos de entrenamiento.

Sobre el uso de conjuntos de datos para el entrenamiento de modelos comerciales privativos, la cuestión es: o bien se están usando datos protegidos por derechos de explotación (copyright), o bien se están usando datos no protegidos por esos derechos (copyleft).

Si se están usando datos protegidos por derechos de explotación (datos que tienen copyright), no termina de estar claro si eso es una infracción que vulnera derechos (los de explotación). Desde la ética hacker siempre se ha criticado muy duramente la protección de obras artísticas o intelectuales con derechos de explotación, porque eso es una limitación a la libre circulación de las ideas y del conocimiento. Por eso, rechinaría ahora hacer valer los derechos de explotación, aunque sea para ir en contra de las grandes corporaciones. El desequilibrio económico y de poder entre una persona que trabaja como fotógrafa, diseñadora, traductora o columnista (por citar solo algunas profesiones) y las grandes corporaciones es tan enorme, la precarización laboral es tan galopante, y la expropiación y reprivatización del conocimiento que están haciendo las grandes corporaciones es tan bestia que la defensa de los derechos de explotación resurge una y otra vez como posible escudo de defensa. Ahí la crítica hacker sigue teniendo actualidad, pues los derechos de explotación no son un camino real para acabar con la precariedad. Las alianzas

entre lo hacker y lo precario abren terrenos fértiles para actualizar el discurso y las luchas.

Por otra parte, si se están usando datos licenciados como copyleft, el debate es otro distinto. Es seguro que se están usando no solo porque están accesibles, sino porque a veces son los únicos disponibles. Hay chatbots específicos para generar código en lenguajes de programación. Esos chatbots han sido entrenados con código de programas de ordenador y el código disponible es el del software libre, porque el software privativo justamente lo que hace es no publicar el código. Por ejemplo Deep TabNine, un asistente para escribir código, ha usado unos dos millones de archivos de GitHub. Se desprende que esos millones de archivos habrán sido obtenidos de los repositorios públicos de GitHub, que son los que utilizan los proyectos de software libre.

GitHub es una plataforma, propiedad de Microsoft, que usan las programadoras para alojar el código y desarrollarlo colaborativamente. Ofrece repositorios públicos y privados. Ha desarrollado su propio asistente para escribir código: GitHub Copilot. Juega con ventaja respecto a Deep TabNine, porque es de suponer que para el entrenamiento habrá usado el código tanto de los repositorios públicos como de los privados. Y de hecho entre las programadoras hay dudas sobre la privacidad de los repositorios privados, cuando se usa Copilot como asistente para aumentar la productividad.

Volviendo a los datos copyleft, el movimiento hacker siempre ha defendido el derecho al conocimiento. En principio, por tanto, sería reacio a imponer restricciones a su acceso y a su uso. Pero también ha luchado siempre contra la privatización, así que garantizar que los modelos de negocio de los chatbots conversacionales no suponen una reprivatización de lo libre es un conflicto abierto.

La reprivatización, el monopolio sobre el conocimiento, no es la única amenaza. También hay un riesgo de colapso. Ahora los modelos se están entrenando con datos generados por humanos, pero, si los humanos masivamente utilizan los modelos para generar nuevos datos, los futuros modelos se entrenarán con datos generados por los anteriores modelos. Teniendo en cuenta que se trata de modelos estadísticos, que dan por bueno lo más frecuente, el acceso a lo minoritario, a lo singular, a lo heterodoxo, a lo inconveniente o a lo raro va a ser, sencillamente, imposible. Solo estarán masivamente accesibles las ideas *mainstream*. ¿Habrá que armar nuevos circuitos para que circule el pensamiento disidente, nuevo, inédito, marginal?

Lo hacker no está muerto. En estas breves líneas he hablado solo de lo cercano, pero hay mucho *hacking* distribuido recorriendo el mundo con mapas diversos. Quizás no todos los proyectos hacker son exactamente congruentes entre sí. Lógico. Pero hay muchas comunidades de desarrolladoras que han contribuido a inventar la inteligencia artificial y que están por la labor de democratizar los sistemas, de que haya modelos públicos, datos transparentes y formas de gobernanza responsables, equitativas y seguras.

Ya hemos hablado de Stockfish, Leela Chess Zero o Apertium. Son solo chispas. La cantidad de contribuciones hacker a la inteligencia artificial es de calado: lenguajes como Phython, librerías como TensorFlow, modelos como Mistral o Stable Diffusion, grupos de investigación como EleutherAI… Hay muchos espacios de cooperación y muchas ideas circulando. En la construcción de la inteligencia artificial la industria no es el único agente.

Lo hacker puede parecer desaparecido, pero es solo porque se ha socializado. Ha explotado, se ha propagado y ya no cabe en un pequeño sótano de un vetusto hacklab. Los conflictos que las tecnologías y las ideologías plantean no son, solo, asuntos del código, asuntos de las programadoras o de las hackers de la línea de comandos.

La aureola heroica del hacker como *cowboy* solitario y errante que va por el mundo digital defendiendo su propia ley no tiene caso. La lucha contra la discriminación algorítmica involucra a académicas y a periodistas, además de a activistas por los derechos humanos. La lucha contra la tecnoprecariedad, a grupos de trabajadoras, sindicalistas y abogadas laboralistas. La lucha por la transparencia y el derecho a auditar, a organizaciones que puedan sostener largas litigaciones estratégicas. La denuncia de las formas de violencia, a las feministas y a los colectivos concernidos. La defensa medioambiental, a ecologistas. El planteamiento de las preguntas existenciales, a filósofas. Detrás del contrapoder capaz de influir en las regulaciones, capaz de defender lo público, de construir modos humanos de convivir con las máquinas…, hay muchos grupos, muchas activistas, muchas resistencias que no son asimilables al estereotipo hacker, pero sin las cuales lo hacker no tiene nada que hacer. Y, al mismo tiempo, hay muchas hackers de la línea de comandos deseando experimentar.

Toda esta socialización es muy bienvenida. Ya era hora. Pero no supone que lo hacker haya perdido su papel específico. Lo sigue teniendo, porque lo hacker entiende, desde lo tecnológico, que las meras

soluciones técnicas a problemas sociales son cortinas de humo que ocultan las asimetrías de poder. Lo hacker lucha por la redistribución del poder, lucha contra su concentración. Y lucha en colaboración con las máquinas, a golpe de la potencia y de la belleza del código.

Nosotras

Este libro llega a su fin y es el momento de exponer mi propio punto de vista. He tratado algunos de los conflictos principales, pero soy consciente de que a otros, como por ejemplo al medioambiental, no les he dedicado suficiente atención. Tampoco me he parado a elaborar cómo la inteligencia artificial está entrando en la enseñanza, en la sanidad o en la investigación en diferentes ámbitos científicos y técnicos. Esto no obedece a una jerarquía de conflictos o reivindicaciones, sino a las limitaciones de mi capacidad de investigación. Me he atrevido a hablar más de lo que más conozco.

Después de haber esbozado algunos de los naipes imprescindibles en una baraja que nos dé juego para relacionarnos con la inteligencia artificial, ahora voy a señalar algunas bazas posibles, a identificar algunas situaciones jugables. Estas situaciones no están ordenadas por gravedad o prioridades. Es solo mi mapa intuitivo de puntos calientes.

Lo primero sería dejar de relacionarnos con la tecnología de modo instrumental. Cada vez que alguien se pregunta para qué sirve esto o aquello, miles de posibilidades mueren y algo humano se quiebra. La cuestión no es si hay usos buenos o usos malos. La cuestión es dejar de usar las cosas, las personas, los animales, las plantas, las bacterias, las máquinas… El mundo no está ahí para usarlo. Está para convivirlo. Y convivirlo es establecer relaciones de cooperación con los seres vivos y también con los seres artificiales. No tener máquinas sirvientas. Tener máquinas amigas. Elegir bien las máquinas que me rodean, las que voy a cuidar y de las que me voy a responsabilizar. Aunque, claro, no podemos aspirar a tener una relación con los algoritmos

distinta a la que tenemos con la lavadora, el frigorífico, el microondas, el ascensor, el teléfono móvil o el coche… salvo que asumamos un proceso de transformación profundo.

Muchas de esas máquinas son cerradas. Muchos de esos algoritmos también. No se prestan a que nos relacionemos con ellos con una tonalidad afectiva amorosa. Son expresiones de lo humano y condensan relaciones de poder. Pero hay otras máquinas. Tiene que haberlas y, si no las hay, hay que construirlas. El software libre y sus comunidades, por ejemplo. Software amigo, comunidades que, en mayor o menor medida, se llevan bazas. Software que se cuida y se defiende, como cuando la Fundación Mozilla se planteó abandonar el cliente de correo electrónico Thunderbird y finalmente dio marcha atrás. Como cuando en Wikipedia se discutió si incluir publicidad para conseguir dinero y se decidió no hacerlo. Como cuando la empresa Sun Microsystems abandonó el desarrollo de su *suite* ofimática y la comunidad se activó para que LibreOffice siga vivo.

Simondon dice que es difícil liberarse transfiriendo la esclavitud a otros seres, sean humanos, animales o máquinas. Reinar sobre un pueblo de máquinas que convierte en siervo al mundo entero sigue siendo reinar, y todo reino supone la aceptación de esquemas de servidumbre. La insubordinación a esos esquemas plantea el desafío de mantener con las máquinas una relación social. Y eso abre cuestiones existenciales, porque es contradictorio defender un humanismo en el que el humano está al frente de las máquinas y a la vez construir máquinas cuya máxima perfección consiste en poder delegar en ellas completamente la decisión humana, cuya perfección radica en que podamos desentendernos. (Y esto no ocurre solo con los algoritmos. Ocurre con todos los automatismos que nos rodean, como el frigorífico, las luces que se encienden solas o las *playlist* de Spotify).

En el frenesí de la digitalización, cada vez es más imposible introducir contenidos formativos que se salgan de lo instrumental. Solo se pueden enseñar cosas prácticas, lo cual es otro modo de decir cosas que aumenten la productividad. Todo lo que no aumente la productividad, fuera. El deseo de saber por saber, el placer de aprender por aprender, fuera. No sirve para nada. En todo caso, se reserva para el arte o la cultura, pero en tecnologías todo ha de ser útil. Basta con mirar el marco europeo de competencias digitales DIGCOMP, que es el estándar para la impartición de formación en tecnologías digitales para la ciudadanía.

Hace diez años yo impartía talleres de programación creativa dirigidos a mujeres que simplemente querían entender qué es eso de programar. Ahora es prácticamente imposible encontrar financiación para ese tipo de actividades. Enseñar cosas que se salgan del guion es perder el tiempo. Y nadie tiene tiempo que perder. Ni siquiera si eres una niña de ocho años que hace robótica. Ni siquiera si eres una jubilada que tiene ansia de saber.

En un sistema económico y social que ha encumbrado la utilidad como el mayor de los valores, que da por hecho que la relación instrumental es la que va de suyo, desplazar la utilidad y en su lugar colocar la pregunta por el sentido es toda una revolución. La utilidad y el sentido son cosas diferentes. La utilidad me viene dada. El sentido me lo he de inventar yo.

Partir de la utilidad no lleva muy lejos. Mientras estoy usando un martillo, no puedo hacerme preguntas sobre el martillo. Bastante tengo con dar en el clavo y no en el dedo. Para hacerse preguntas hay que separarse de lo útil. Por ejemplo, se puede hacer un taller con personas jóvenes para que se hagan preguntas sobre el móvil teniendo el móvil en la mano. Cierto. Pero quizás las preguntas sean mejores si el taller consiste en aparcar el móvil durante veinticuatro horas y luego ver qué ha pasado. No cómo funciona el móvil, sino cómo (no) funciono yo sin el móvil. Para qué sirve el móvil ya lo sabemos. Pero el sentido de pasar tiempo con móvil o sin él, eso es una creación.

Cuando la tecnología se ofrece como algo en lo que poder delegar, tan importante como lo que hace la máquina es lo que deja de hacer el humano. Si delego mi desplazamiento en el ascensor, dejo de subir escaleras. Mi físico se debilita. Con la delegación en los automatismos algorítmicos, las escaleras que vamos a dejar de subir son las de la toma de decisiones y la responsabilidad sobre ellas. Dejaremos de practicar la toma de decisiones. ¿Se debilitará mi moral? ¿Dónde quedará la responsabilidad? ¿Dónde la libertad?

Decidir por dónde voy de un lugar a otro o qué frase pongo en un texto puede parecer un asunto menor. Y quizás lo sea. Pero hay que pensar qué consecuencias puede tener, en la libertad y en la responsabilidad, el hecho de desacostumbrarnos a tomar decisiones. Por más que se presenten como auxiliares, en la práctica muchas máquinas las están tomando por nosotras. Automatismos como Lavener, que tal vez está utilizando el ejército israelí. O automatismos como VioGén.

Lo fácil es endosar la responsabilidad de la toma de decisiones a las desarrolladoras, a las científicas de datos, a las profesionales de la programación. Pero esto es como decir que la doctora es la única que debe decidir sobre cuándo aplicar la eutanasia. Es echar balones fuera. Es delegar en las expertas cuestiones que están muy lejos de ser objetivables y que responden, deben responder, a valoraciones, a valores construidos y compartidos socialmente. A alianzas tecnosociales.

En el río revuelto por lavarse las manos y encasquetar responsabilidades, la industria encuentra argumentos contra el software libre: si la programadora es la responsable del uso, entonces no podemos hacer software libre, porque el software libre puede ser utilizado por cualquiera para lo que quiera. Si queremos controlar el uso, entonces hay que hacer software privativo. Si la responsabilidad se delega en lo puro técnico, entonces lo social se queda sin bazas que jugar.

Lavender queda muy lejos, pero VioGén queda muy cerca. Con VioGén hay bazas. Pensar en cómo ese algoritmo predictivo debería ensamblarse en el sistema de lucha contra la violencia de género, en cómo podría o debería ser el algoritmo en sí y también en cuál sería un buen contexto de uso no es fácil, pero está al alcance de un movimiento feminista que está organizado y activo y que puede hacer alianzas con científicas de datos para jugar las cartas técnicas y políticas.

María Salguero, ingeniera en geofísica, es una activista de datos que se ha volcado en documentar los feminicidios en México. Ha creado una base de datos que va actualizando con informaciones que se publican en los medios de comunicación. Para cada feminicidio registra trescientos datos, que son los que considera necesarios para recopilar toda la información sobre el asesinato y su contexto. ¡Trescientos datos por asesinato! (Viogén evalúa el riesgo de reincidencia con treinta y cinco preguntas). Su web se llama *Yo te nombro. El mapa de los feminicidios en México*, y ahí publica un mapa geolocalizado para hacer visibles los lugares donde las están matando, encontrar patrones, reforzar los argumentos sobre el problema, georreferenciar la ayuda, promover la prevención e intentar evitar los feminicidios. Es uno de los muchos proyectos, personales o colaborativos, para hacer ciencia de datos feminista.

La ciencia de datos feminista es un concepto acuñado, entre otras, por Catherine D'Ignazio y Lauren Klein. En su publicación *Data Feminism*, traducida y disponible en Internet, exponen los criterios que

definirían el concepto: examinar el poder, considerar la emoción y el cuerpo; valorar múltiples formas de conocimiento, asumir que los datos no son neutrales ni objetivos, pues son producto de relaciones sociales desiguales… Un activismo científico que se pregunta qué información debe convertirse en datos antes de que se pueda confiar en ella.

Para estas investigadoras, el feminismo interseccional no trata solo de mujeres. Ni siquiera de género. Tiene que ver con el poder: con quién lo tiene y quién no. Y en un mundo en el que los datos son poder, el feminismo de datos puede ayudar a comprender cómo desafiar y cambiar el sistema. La tecnología tiene potencial para hacer el bien. Hay que ver qué puede aportar el feminismo al *big data* y a la inteligencia artificial. Y, en este camino de experimentación, el algoritmo VioGén está relativamente al alcance de la mano. Hay recorrido posible. Hay bazas.

Creo que también las hay en el asunto de la soledad y los cuidados en la vejez. La relación entre soledad y vejez parece natural, pero no lo es. ¿Por qué hay tanta soledad al final de la vida? ¿Sabemos cómo queremos vivir la vejez? ¿Y cómo queremos vivir antes de que llegue?

El movimiento de vida independiente surgió como parte de los movimientos por los derechos civiles. Protagonizado por personas con diversidad funcional, defiende que las diferencias físicas, intelectuales o sensoriales no pueden ser un argumento para que las personas pierdan el derecho a ejercer el control sobre sus vidas. Reivindican, básicamente, no ser institucionalizadas, contar con una persona que trabaje como su asistente personal y disponer de ayudas tecnológicas. Su lema es «nada sobre nosotros/as sin nosotros/as», algo que las personas ya muy mayores quizás no pueden defender por sí mismas.

Obvio que no es lo mismo diversidad funcional que vejez, pero es un movimiento de largo recorrido que trabaja por definir qué es una vida independiente. A partir de una realidad humana diversa, plantea en primera persona cuestiones que pueden servir de inspiración o de orientación, como la denuncia de las condiciones de trabajo de lo que llaman el ejército de esclavas de los cuidados. O la defensa de asistencias técnicas que dejen espacio abierto a las necesidades, gustos y preferencias personales.

En un post en la web del Foro de Vida Independiente, Adolf D. Ratzka publica: «Un aparato asistencial puede disminuir o aumentar

las limitaciones. Conservé mi silla de ruedas durante veintidós años porque era simple y liviana pero resistente, ideal para viajes y aventuras. Una persona podía ayudarme a subir un bordillo alto con facilidad, dos personas podían ayudarme a subir una escalera. Este tipo de silla ya no está disponible: aparentemente las grandes ruedas traseras son consideradas peligrosas para mi seguridad. Las sillas eléctricas con pequeñas ruedas traseras me confinan a ambientes accesibles como centros comerciales e instituciones. De este modo, lo que otras personas asumen que es mejor para mí limita mi movilidad, me niega la dignidad de tomar riesgos y cuestiona mi habilidad de actuar buscando mis mejores intereses». Está exigiendo poder codiseñar la máquina con la que acoplarse. No es lo mismo una máquina que otra.

En una sociedad en la que hay que gestionar la vida propia como un emprendimiento, todo lo que evidencie la fragilidad pincha el globo del «sí se puede». Hay cosas que no se pueden. Pero toparse con el límite causa tristeza cuando el límite está despojado de potencia. Ver que alguien se va quedando sin fuerzas es pesaroso. Desde la lógica de que en la vida no hay tiempo que perder, la vejez es una pérdida total. Es una avería, un fallo en el sistema. Es una errata que no dice nada, que no aporta nada. Un error de diseño. Y ahí está el transhumanismo para repararlo.

Para la industria, la vejez es un excelente mercado de consumo. Pero, más allá de esta lógica de mercado, hay que comprender no ya lo que las viejas farfullan, sino lo que la vejez, como muestra de los límites de lo humano, dice a la sociedad entera. No puede ser que la vejez, como realidad humana, no tenga potencia. Es imposible. Esa potencia tiene que estar y hay que encontrarla. ¿Qué le dice la vejez a lo humano? ¿Y qué responde la ideología de la inteligencia artificial?

Entre instituciones, ejércitos de esclavas de los cuidados, soledad, asistencias técnicas, *cohousing* sénior, robots sociales y familias desbordadas hay recorrido, y necesidad, para la elaboración y la experimentación. No solo cuidados que palíen la soledad. No solo la presencia tranquilizadora de alguien que me hace compañía y de quien espero pasivamente que llene lo que está vacío. También escuchar a una vejez que clama a voz en grito su debilidad y saber descubrir la potencia de esa endeblez.

La ideología de la inteligencia artificial considera que la vejez o la enfermedad son errores de diseño en lo humano y que se pueden

reparar. Si se pueden reparar, ¿por qué no habría que hacerlo? Llevado al extremo, el transhumanismo parece una ida de olla. Pero lo cierto es que en la mayoría de los países europeos la legislación permite el aborto cuando se puede predecir con probabilidad o certeza que el feto nacerá con un defecto o enfermedad grave incurable. La caracterización de lo que es enfermedad y lo que es diversidad es difusa. El síndrome de Down, por ejemplo, no es una enfermedad. Es una variación genética en el ADN. El hecho es que cada vez nacen menos personas con este síndrome. Cada vez hay menos embarazadas que los den a luz. Esta cuestión involucra libertad, inclusión y diversidad: decisiones aparentemente personales, pero que en realidad son un asunto colectivo. ¿Qué condiciones de justicia y equidad favorecen la diversidad? Estamos aceptando vivir en una sociedad en la que no hay lugar para lo humano ¿disminuido?

Esta cuestión no se resuelve con algoritmos, pero los algoritmos vienen envueltos en ideas que impregnan, que tiñen todos los asuntos. Las ideas que se aplican en un ámbito se expanden fácilmente hacia otros que, aparentemente, no tienen nada que ver. Y está entrando una idea con la que habría que saber qué hacer: la idea de que el cuerpo no aporta nada. Si algo podemos agradecer a la inteligencia artificial, es que plantee con tanta crudeza el asunto del cuerpo con todas sus derivadas: la vejez, la enfermedad, la diversidad, la elección, el rechazo, la legitimidad de diseñarlo y domeñarlo… Para pensar el transhumanismo hay bazas.

En esta brevísima enumeración de jugadas posibles termino con la del trabajo y el salario social. Detrás de un modelo de lenguaje, como por ejemplo ChatGPT, producido con lógicas extractivistas, hay mucho trabajo invisibilizado, mal pagado e incluso gratuito. Pero no es la única manera de hacer las cosas. Si necesitamos datos, podemos generarlos, etiquetarlos y hacer todo lo que haya que hacer con ellos sin que forzosamente tenga que haber una explotación salvaje del trabajo. Igual que tenemos un Wikipedia, podemos tener conjuntos de datos en términos de propiedad colectiva o pública gestionados como bienes comunes.

El conocimiento se produce colectivamente, por cooperación. No solo por cooperación voluntaria y consciente. Se produce por cooperación a secas. Simplemente viviendo ya contribuimos a él. El inventor aislado al que le vienen a la cabeza ideas geniales es, sencillamente, un mito imposible. El conocimiento se expropia de quienes

lo producimos, se privatiza y luego se nos ofrece como servicio. Pero en ese tránsito hay fugas, hay pérdidas, hay robos.

Internet se concibió y se construyó como una red abierta, libre y neutral. Pero esta apertura tal vez se esté cerrando. Antes todas las comunidades de programación organizaban todos sus debates en foros públicos. Ahora sigue habiendo foros públicos, pero también hay muchos debates que tienen lugar en grupos de Telegram. Los grupos de Telegram no son públicos. Gran parte de los debates mundiales se está privatizando. Esto recentraliza un conocimiento que antes era distribuido. Telegram sabe de todos los debates mundiales. Nosotras ya no. Telegram podrá entrenar inteligencias artificiales con esos debates. Nosotras no. Es la fragmentación de las comunidades. Es su muerte. En grupos privados puede haber mayor afinidad y menos fricciones, pero en lo que respecta a redistribuir el conocimiento es pan para hoy y hambre para mañana.

Elon Musk dijo que iba a eliminar los *hashtags* de su red social. En el antiguo Twitter, y ahora en X, la gente utiliza los *hashtags* para vincular unos contenidos con otros y eso permite a las lectoras llegar a contenido nuevo relacionado. Si se eliminan, la información sobre las relaciones desaparece. El caso es que esto puede hacerlo cuando quiera, que para eso la red es suya. Mejor dicho, tal vez lo que haga será invisibilizar los *hashtags*, pero no borrarlos. Así solo él tendría acceso a esa información sobre los vínculos.

Esa información que ya no veremos, pero que es nuestra, porque la hemos generado a pico y pala, puede alimentar un modelo de lenguaje que después se nos ofrecerá como servicio y del que con la boca abierta diremos que sabe más que nosotras. Todo el conocimiento que corre por los grupos de Whatsapp es visible, en su conjunto, solo para Whatsapp. Alimentará modelos de lenguaje a los que preguntaremos cómo son las cosas. Cómo son nuestras cosas.

El salario social es el reconocimiento de que lo social trabaja y produce. Vivimos rodeadas de automatismos y al mismo tiempo cada vez trabajamos más. ¿Cómo es posible? Las luchas por la renta, por el reparto equitativo de los logros del conocimiento, tienen que ver con las tecnologías. Queremos y merecemos buenas tecnologías. Queremos y merecemos riqueza vital.

En definitiva, jugar en la inteligencia artificial no es solo resistirse y luchar para que se implanten contenciones éticas. No es solo luchar por las regulaciones y denunciar las relaciones de poder. También

es disputar políticas digitales, defender y construir lo público y lo común. Es aceptar la fricción, ensuciarse las manos. Es desmontar la máquina y volverla a montar. Generar dinámicas de investigación, de construcción, de taller. Dar sentidos y quitar sentidos. Articular de modo conflictual los saberes experimentales y los tecnológicos. Sentir alegría por la potencia de actuar, de crear.

Termino de escribir estas líneas en noviembre de 2024 mientras miro fotografías de la manifestación en Valencia que exige la dimisión del *president* Mazón por su criminal no gestión de la DANA. En una pancarta puede leerse: «Estamos buscando tantas respuestas que olvidamos las preguntas».

Las preguntas son los posibles del porvenir que está ya en el presente. Definir qué es lo humano no está dado de una vez por todas. Cada época debe dictar su definición, elaborarla, perfilarla. La inteligencia artificial abre preguntas de mucho calado sobre la tecnología, que nos hizo humanos, sobre las relaciones con lo otro y con los otros, y está en la encrucijada de todo lo que importa. También abre el duelo por la prepotente visión que el hombre tenía de sí mismo.

Ojalá podamos decir que haber vivido la época en la que nos topamos con la inteligencia artificial fue interesante. Ojalá podamos decir que mereció la pena.

Idear

Carta a Charo

Querida Charo, ¿cómo vais?

Yo ya he terminado de escribir el libro. He metido todas las sugerencias que me diste.

¡No veas lo que he disfrutado! Aunque también te digo que, cuanto más se supone que sé, más inseguridades tengo. Es como el mundo al revés, como si el conocimiento fuera una carrera hacia las dudas, hacia las preguntas sin respuesta.

¿Qué hacer con todo esto? Pues experimentar, ¿no?

He pensado que cuando terminemos lo que estamos investigando ahora con ChatGPT quizás podríamos recuperar aquello que hacíamos en los talleres de programación creativa y poner en marcha algún proyecto que nos sirva para aprender. No tanto adiestrarnos en usar inteligencias artificiales ya empaquetadas, sino abrir camino para ponernos nosotras a crear.

A lo mejor, si tuneamos el proyecto como algo de capacitación para el empleo, podríamos conseguir financiación.

Bueno, Charo, un día de estos comemos juntas y seguro que nos vienen las ideas.

Muchas ganas de tramar algo.

Besos.

Despedida

He escrito este libro a la antigua usanza, poniendo una palabra después de la otra. No he usado ChatGPT, pero sí buscadores y traductores automáticos. Aunque las palabras las he puesto yo, las ideas son colectivas. No hay ni una sola idea original. Todo está copiado de alguien, inspirado en algo, reformulado a partir de otra cosa. Solo por convención lo firmo como si fuera mío.

Margarita Padilla García
mpadilla@sindominio.net