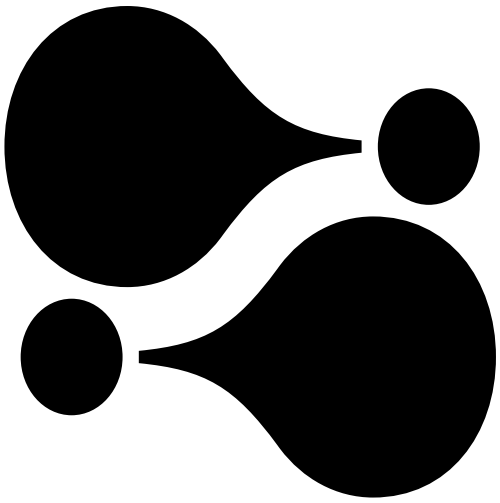
Algorismes responsables?

Un matí amb Louise Amoore

Propostes per treballar a l'aula



CCCB Centre de Cultura Contemporània de Barcelona

Índex

Dossier elaborat per Roger Manau amb el suport de l'equip del CCCB.

L'algorisme, classes d'algorismes i capacitats que tenen

Text adaptat de l'article de José Antonio Estévez «La algorimización en el mundo del capitalismo de la vigilancia» a *Oxímora: Revista Internacional de Ética y Política*, núm. 20, gener-juny de 2022, pàg. 1-37. ISSN 2014-7708.

—

L'algorisme, classes d'algorismes i capacitats que tenen

Què és un algorisme? És una pregunta que rarament es formula. En general, es dona per descomptat que tothom en sap la resposta. Però quan algú té l'atreviment de plantejar explícitament la qüestió en un fòrum en què es pot haver estat parlant durant hores o fins i tot dies de la regulació de l'ús dels algorismes, resulta que els tinguts per experts no es donen per al·ludits o fugen d'estudi de manera més o menys afortunada.

Després d'indagar-hi durant bastant de temps sense haver reeixit a trobar-ne una formulació universalment acceptada, la concepció que em sembla més clara, general i convincent és que **un algorisme és una seqüència de passos o conjunt d'instruccions que s'han de seguir per resoldre un problema determinat.** Així doncs, una recepta de cuina seria un algorisme que ens permet solucionar la qüestió de com fer, per exemple, una bona truita de patates. D'acord amb aquesta concepció, els algorismes són molt més antics que la revolució digital. Pitàgores en va dissenyar alguns de molt enginyosos per resoldre problemes geomètrics. Però quan avui dia parlem d'algorismes ens referim generalment als de caràcter computacional, és a dir, a **seqüències d'instruccions que poden ser compreses i executades per un ordinador o per un dispositiu complex que té entre els seus components un computador.**

Algorismes deductius i inductius

Cada algorisme, o almenys cadascun dels tipus d'algorismes que hi ha actualment, és un món en ell mateix. Per aquest motiu, és difícil fer gaires afirmacions que es puguin aplicar a tots i cadascun sense matisos. No podem explorar aquí tots aquests universos, però, pel que fa a l'anàlisi dels problemes eticopolítics que plantegen, com a mínim cal diferenciar dues grans classes d'algorismes, que anomenaré *deductius* i *inductius*.

Els algorismes deductius estan constituïts per sèries finites d'instruccions o passos elaborats sobre la base d'un model matemàtic que serveix per seleccionar les característiques i les variables que es consideraran rellevants. Estan compostos de sèries de regles ordenades que tenen l'estructura sintàctica pròpia de les sentències condicionals. És a dir: «Si es dona la situació A, aleshores fes l'operació B.» La seqüència de passos que cal seguir per solucionar el problema la dissenya de principi a fi un programador. L'ordinador es limita a seguir les instruccions de manera anàloga a com ho fa amb bona part del programari que fem servir habitualment a casa o a la feina.

En canvi, els algorismes inductius són ben diferents. Louise Amoore, en el seu llibre *Cloud ethics* (2020), se centra particularment en els algorismes utilitzats en un tipus de màquines virtuals dotades d'intel·ligència artificial denominades *xarxes neuronals*. Aquests mecanismes es diferencien dels que fan servir algorismes deductius en dos aspectes crucials. D'una banda, són **capaços d'aprendre i de reprogramar-se autònomament. I, de l'altra, els resultats a què arriben poden ser imprevisibles i inexplicables fins i tot per als programadors que els han dissenyat.** L'autora insisteix sovint al llibre en la diferència entre aquests dos tipus d'algorismes: «La representació dels algorismes com una cadena lògica passa per alt el grau en què **els algorismes es modifiquen a si mateixos en i a través de les seves relacions iteratives no lineals amb les dades d'entrada»** (Amoore, 2020, p. 11).

Què poden fer els algorismes?

Els algorismes són capaços de fer moltes menes d'operacions diferents. Ara ens interessen especialment les seves **capacitats de reconeixement, classificació, establiment de correlacions estadístiques, aprenentatge i reprogramació.** Els algorismes poden reconèixer objectes o persones a partir de dades digitalitzades proporcionades per diversos tipus de sensors o dipositats en diverses classes de bancs.

Un algorisme no només és capaç de reconèixer una cara en una escena o una imatge, sinó també d'identificar a quina persona correspon. És el que fan els sistemes de reconeixement facial que fa servir la policia en diversos països i dels quals hem parlat més amunt. Aquests sistemes utilitzen sofisticades càmeres de vigilància i bancs de dades —en els quals s'han recopilat nombroses imatges de persones d'interès, si no tota la població— per identificar qui participa en una protesta o qui deambula per un aeroport. Els algorismes poden determinar si un objecte o una persona pertanyen a una determinada classe. Ho fan examinant si té els atributs que defineixen el conjunt de què es tracti. Per exemple, són capaços de decidir si les característiques d'un missatge de correu electrònic permeten o aconsellen classificar-lo com a correu brossa. Perquè un algorisme sigui capaç de reconèixer i classificar persones i coses, ha de tenir la capacitat d'aprendre. Tots els algorismes es programen, però alguns també s'han d'entrenar. Se'ls en pot ensenyar de diverses maneres. Per exemple, utilitzant les dades etiquetades que hem esmentat perquè aprenguin a classificar coses o persones.

Els algorismes que han de fer operacions de reconeixement visual s'entrenen mostrantlos una gran quantitat d'imatges. Si han de reconèixer números escrits a mà, es fa servir una base d'imatges de dígits manuscrits (a internet se'n troben) perquè aprenguin a identificar patrons. Això els permetrà reconèixer si el número escrit al paper és un 1, un 7 o un 9. Si es tracta d'un sistema de reconeixement facial, l'entrenament consistirà a mostrar imatges de rostres humans extretes de documents oficials o de les xarxes socials. Un cop enllestit

el període d'instrucció, alguns tipus d'algorismes són capaços de continuar aprenent per compte propi. Mentre operen, detecten noves característiques rellevants i perfeccionen les seves classificacions afinant, per exemple, la seva capacitat d'identificar quins missatges són correu brossa. També creen nous conjunts de coses o persones identificanthi atributs nous que resulten adequats per classificar-les, a fi de resoldre els problemes per als quals han d'oferir solucions. Els algorismes poden establir correlacions estadístiques entre conjunts de persones o coses. És a dir, poden calcular la probabilitat que, si una persona pertany al conjunt A, també formi part del conjunt B. Però també poden descobrir per si mateixos correspondències inèdites i sorprenents sense necessitat de partir d'una hipòtesi. Aquesta capacitat resulta especialment útil en les anàlisis de grans bancs de dades, com els designats com a *big data*.

—

Preguntes:

1. Després de llegir el text, com definiries què és un algorisme?

2. Quines són les cinc característiques principals dels algorismes, segons el text?

3. En parelles, feu una llista d'algorismes presents a la vostra vida quotidiana. Després, comenteu com influeixen en la vostra presa de decisions. Per exemple, podeu pensar en un recomanador de pel·lícules d'una plataforma (Netflix, HBO...), o l'assistent virtual d'una pàgina web de compres.

4. Amoore assenyala que hi ha algorismes inductius i algorismes deductius. Quina diferència hi ha entre aquests tipus d'algorismes? Quines capacitats especials tenen els algorismes inductius?

5. Feu un debat a classe sobre els avantatges i els perills dels algorismes en la societat actual.

Com els algorismes perpetuen la desigualtat a Espanya

Article publicat a *La Vanguardia*[1] el 26 d'octubre de 2021.

1
www.lavanguardia.com/tecnologia/20211026/7814332/como-algoritmos-perpetuan-desigualdad-espana.html

—

Com els algorismes perpetuen la desigualtat a Espanya

Itziar P., una psicòloga de Castelló, va anar un dissabte de febrer de 2018, a primera hora del matí, a denunciar que el seu marit, Ricardo Carrascosa, l'amenaçava. Els agents van agafar la denúncia i van passar les seves respostes per l'algorisme Viogen, que ajuda la policia a estimar el risc de reincidència per violència de gènere. El programa va qualificar la situació de «lleu», i el jutge va basar-se en aquest document per denegar l'ordre d'allunyament que la dona sol·licitava. Set mesos més tard, Ricardo va assassinar brutalment les filles que tenien en comú, Nerea i Martina, de sis i dos anys. El programa havia fallat.

Els algorismes —les fórmules matemàtiques darrere d'un ordinador— automatitzen cada vegada més decisions que afecten el dia a dia de la ciutadania. Però, si a la societat que els programa i els aplica hi ha racisme, masclisme i classisme, els algorismes no estaran exempts d'aquests biaixos. Cedir el poder de decisió a les màquines confiant en el seu criteri de *neutralitat* també implica continuar enfortint, de manera invisible, les desigualtats socials.

Malgrat que estats i empreses els utilitzen de manera recurrent, l'opacitat és el denominador comú en aquesta àrea. Pràcticament és missió impossible que un ciutadà sàpiga quan se l'està sotmetent al cribratge d'un algorisme, quins factors té en compte i si hi ha cap criteri ètic que en controli el funcionament. En aquest sentit, l'entitat Eticas Foundation ha volgut posar-hi llum i ha creat l'Observatori d'Algorismes amb Impacte Social (OASI), que aplega

i analitza l'impacte social de més de cinquanta algorismes a tot el món. En el cas de l'Administració espanyola, analitza tres sistemes que s'apliquen de manera pràcticament desconeguda per a la majoria de la població, i un quart que s'aplica només a Catalunya: tots podrien arribar a perpetuar la discriminació socioeconòmica i la vigilància de l'Estat.

Amb l'objectiu d'estalviar temps i evitar fraus, la Policia Nacional espanyola va implementar el 2018 un sistema conegut com a Veripol per identificar les denúncies falses interposades per la ciutadania relatives a robatoris amb violència, intimidació i furts. Segons les dades cedides pel Ministeri de l'Interior a Algorithm Watch, una organització sense ànim de lucre que observa i analitza les preses de decisió automatitzades, s'ha utilitzat aquest instrument en aproximadament 84.000 denúncies. Inicialment, aquest algorisme va ser entrenat amb 1.122 denúncies registrades l'any 2015 a Espanya: 534 hi figuraven com a veritables i 588 com a falses. Però les denúncies en si tampoc no són objectives, perquè havien estat redactades per un agent amb una visió pròpia de la realitat. En aquest sentit, Karma Peiró, periodista i investigadora especialitzada en l'estudi dels algorismes, explica que és precisament durant l'entrenament quan, de manera involuntària, s'hi incorporen biaixos: «Els ensinistrem amb dades del passat i d'acord amb patrons que és possible que hagin canviat. A més, les han recollit humans i, per tant, hi incorporen biaixos que estan tan normalitzats que s'invisibilitzen.»

El Servicio Público de Empleo Estatal (SEPE) també disposa del seu propi algorisme, que des de l'any 2012 automatitza l'adjudicació de prestacions a cada desocupat. Des que es va implementar, la quantitat de persones que reben un subsidi s'ha reduït més d'un 50 %. Segons Eticas Foundation, aquesta diferència no només s'atribueix a l'algorisme, sinó també a la millora de la metodologia a l'hora de creuar les bases de dades, que permet reduir errors. En relació amb aquesta qüestió, l'any 2019 els treballadors del servei públic d'ocupació de Suècia —que, com a Espanya, també té un sistema automatitzat de repartiment de prestacions— van revelar

que un frau de l'algorisme havia deixat sense prestació un total de 70.000 desocupats. L'escàndol va obrir un debat nacional sobre l'ús d'aquests procediments per part de l'Estat.

D'altra banda, el Ministeri de l'Interior i la Universitat Autònoma de Madrid també van desenvolupar el seu propi algorisme per identificar missatges d'odi a Twitter. Per bé que és una incògnita si hi ha cap organisme públic que faci servir avui dia aquest programa, Eticas Foundation alerta que l'ús inadequat d'aquesta eina pot promoure el control i la vigilància per part de l'Estat.

A banda dels casos esmentats, a Catalunya també s'utilitza un quart algorisme, que fa servir el Departament de Justícia. S'anomena e-Riscanvi i elabora, des del 2009, un índex de predicció de reincidència criminal que després és tingut en compte a l'hora d'atorgar o no la llibertat condicional a un pres. Després de ser testat al llarg de setze mesos amb 675 presos aleatoris i comparat amb un grup de control de 225 persones, les autoritats catalanes asseguren que els resultats van ser un èxit: la fiabilitat de les prediccions de reincidència va passar del 67 %, amb només criteris clínics humans, al 75 %.

Per a Peiró, el perill no recau tant en els algorismes coneguts com en els que s'apliquen sense el coneixement o consentiment, que tracten dades personals sensibles i que ni tan sols consten encara a la llista d'Eticas Foundation.

És el cas de la terminal d'autobusos més gran d'Espanya, l'estació madrilenya de Méndez Álvaro, que des del 2016 té càmeres de reconeixement facial que comparen automàticament el rostre de cada visitant amb una base de dades de sospitosos de la policia espanyola. Cada any, al voltant de 20 milions de viatgers passen per l'estació i, malgrat això, molt pocs coneixen el tractament que s'està fent de la seva imatge.

I altres exemples: un institut de secundària de Barcelona va instal·lar unes càmeres de reconeixement facial que enviaven automàticament

un SMS als pares quan un alumne no era registrat al matí. En aquest cas, l'Autoritat Catalana de Protecció de Dades hi va intercedir i va retirar el sistema del centre educatiu. O el SAVRY, un instrument de valoració del risc de reincidència en adolescents que van crear uns acadèmics nord-americans l'any 2006 i que l'Administració catalana utilitza actualment.

La intel·ligència artificial no és bona ni dolenta *per se*, però la branca que intenta predir el comportament humà a partir de dades és extremadament sensible. «Sobretot perquè, mentre que aquests sistemes estan a l'ordre del dia, el grau de desconeixement que en té la població és molt elevat», denuncia Peiró.

Les administracions públiques encara han de fer un esforç de transparència per crear un registre d'algorismes que permeti a la ciutadania no només conèixer l'existència dels algorismes, sinó també entendre com operen i conèixer els sistemes de control ètics existents.

—

Preguntes:

1. A l'article es proporcionen exemples d'usos de diversos algorismes per part de les institucions públiques i s'explica que, sovint, s'apliquen de manera esbiaixada. Quines conseqüències pot tenir que un algorisme estigui "esbiaixat"?

2. A l'article se cita la periodista Karma Peiró, especialista en tecnologies de la informació i la comunicació. A partir de les seves declaracions, fes un resum de la seva opinió sobre els biaixos en els algorismes. Si vols aprofundir més, pots consultar aquesta entrevista on també en parla.

3. Per grups, trieu un dels algorismes que es mencionen a l'article i busqueu-ne més informació a internet. Després, responeu les preguntes següents:

   - Quina és la finalitat d'aquest algorisme?
   - Quins beneficis pot tenir?
   - Quins biaixos pot tenir o quins s'han demostrat?
   - De quina manera es podrien corregir els biaixos perquè fos més just?
   - Els ciutadans haurien de tenir accés a la informació que recullen i utilitzen aquests algorismes?

Problemas ético-políticos que plantea la utilización de algoritmos

Fragment adaptat de l'article de José Antonio Estévez «La algorimización en el mundo del capitalismo de la vigilancia» a *Oxímora. Revista Internacional de Ética y Política*, núm. 20, gener-juny de 2022, pàg. 1-37. ISSN 2014-7708.

—

Problemas ético-políticos que plantea la utilización de algoritmos

Los algoritmos realizan ponderaciones al igual que hacen los jueces. En caso de conflicto, los órganos judiciales determinan qué derecho, principio o bien jurídico debe prevalecer. Las premisas y criterios utilizados para la ponderación pueden tener carácter no sólo jurídico, sino también ético o político.

Los argumentos esgrimidos en las sentencias de los tribunales constitucionales para establecer que un derecho fundamental prevalece sobre otro ponen claramente de manifiesto esto último. Louise Amoore hace uso específicamente del término "ponderación" al referirse al modo como los algoritmos inductivos, especialmente las redes neuronales artificiales, razonan: "La disposición de las proposiciones hace que un resultado aparentemente óptimo surja de la ponderación diferencial de los caminos alternativos a través de las capas de un algoritmo (Amoore, 2020, p. 13)".

La diferencia entre los algoritmos y los jueces es que éstos últimos tienen que fundamentar sus sentencias. El juez ha de especificar qué hechos considera efectivamente acaecidos y en base a qué pruebas. Ha de exponer los fundamentos normativos que le han llevado a dictar su fallo en relación con los hechos juzgados. El algoritmo nos da una solución y una probabilidad de éxito que sería

equivalente al “fallo”. Pero el usuario del algoritmo o el destinatario de sus decisiones suelen desconocer cómo ha llegado el algoritmo a esa conclusión. Acceder al “código fuente” no proporciona un conocimiento suficiente de los factores que se han tenido en cuenta ni de las valoraciones que se han llevado a cabo en el caso de los algoritmos que hemos denominado “inductivos”. No nos dirá qué pesos y umbrales de activación han utilizado las neuronas. Desconoceremos de dónde han extraído la información y por qué han seleccionado unos rasgos de los datos en lugar de otros.

El algoritmo nos ofrece una solución a un problema y su probabilidad de éxito (p. ej. un 90%). En su proceso de razonamiento, el algoritmo se encuentra con innumerables bifurcaciones. En diversos momentos puede haber escogido uno u otro camino basándose en una probabilidad menor (p. ej. del 60%). La probabilidad que da a su propuesta final oculta el grado de incertidumbre al que se ha enfrentado a la hora de realizar las opciones previas que finalmente le han conducido a proponer esa solución. Las decisiones de los algoritmos no están, por consiguiente, fundamentadas. No se exponen las premisas, valoraciones y opciones que han conducido a su output. Las decisiones de los jueces son siempre recurribles por los afectados. Las de los algoritmos, no.

Los sesgos discriminatorios constituyen uno de los problemas que más preocupan a quienes tratan el tema de la regulación de los algoritmos.

La distinción entre discriminación directa y discriminación indirecta utilizada en el ámbito jurídico resulta de utilidad para analizar el problema de los sesgos algorítmicos. La discriminación directa se da cuando la ley establece explícitamente tratos diferentes para las personas debido a su raza, género, ideología, religión... sin que exista una justificación razonable para ello. La discriminación indirecta tiene lugar cuando la ley no establece explícitamente distinciones basadas en el género o la raza, por ejemplo, pero el resultado estadístico de su aplicación se traduce en un trato perjudicial para las personas negras o para las mujeres. En el caso de los algoritmos

deductivos, los sesgos discriminatorios pueden evitarse o corregirse técnicamente con relativa facilidad si la discriminación es directa.

En el caso de que el código contenga instrucciones que explícitamente establezcan diferencias de trato no justificadas entre hombres y mujeres o entre personas blancas y negras, será necesario eliminar o modificar dichas reglas reprogramando el algoritmo. Cuando se constata que el algoritmo genera formas de discriminación indirecta, la cosa resulta un poco más complicada. Si su aplicación discrimina estadísticamente a las personas de raza negra o a las mujeres, entonces habrá que descubrir cuál es la instrucción que introduce ese sesgo. El problema puede ser producto, por ejemplo, de un paso del programa que obliga a tener en cuenta el tipo de barrio donde vive una persona a la hora de determinar su grado de solvencia económica. Si se da la circunstancia que en los barrios más pobres habita un mayor número de personas negras, el resultado de la utilización del algoritmo puede ser indirectamente discriminatorio para los afrodescendientes. Pero, en cualquier caso, localizar el defecto puede resultar una tarea bastante laboriosa.

Un algoritmo capaz de aprender autónomamente y de reprogramarse a sí mismo puede también discriminar indirectamente a las personas en función de su raza, género u otras “categorías sospechosas”. El sesgo puede constatarse analizando estadísticamente los resultados que va produciendo. Pero en el caso de estos algoritmos que hemos denominado “inductivos”, es muy difícil evitar o corregir sus tendencias discriminatorias. Los técnicos pueden modificar determinados parámetros del algoritmo, pero no pueden prever con exactitud cómo modificará eso los resultados que éste proporcione.

Amoore cuenta que: “como me explicó un informático, una red neuronal como AlexNet, con seis u ocho capas ocultas, es demasiado compleja para que el propio diseñador del algoritmo pueda delimitar las probabilidades condicionales que aprende. ‘Puedo ajustar la ponderación en esa capa’, explica, ‘y sé que esto cambiará la salida, pero no puedo decir exactamente cómo’. Al igual que con el diseño de

AlexNet, los informáticos trabajan con la naturaleza esencialmente experimental e incógnita del algoritmo (Amoore, 2020, p. 74).

—

Preguntes:

1. Segons el text, quines són les principals diferències entre els jutges i els algorismes a l'hora de prendre decisions? Per què les decisions dels algorismes poden resultar més problemàtiques?

2. Explica breument la diferència entre "discriminació directa" i "discriminació indirecta".

3. Creus que els algorismes inductius, que poden aprendre i reprogramar-se de manera autònoma, s'haurien d'utilitzar en casos que poden afectar els drets i la vida de les persones? Per què? Quines serien les consideracions ètiques a tenir en compte?

4. Per grups, imagineu-vos com hauria de ser un algorisme que decidís l'accés dels estudiants a la universitat. Heu de pensar quins aspectes tindria en compte l'algorisme a l'hora de donar prioritat a uns estudiants o a altres (per exemple: la nota? l'economia familiar?...). Cada grup ha d'intentar que l'algorisme no tingui biaixos i que sigui just i equitatiu en el procés de selecció. Després, poseu-ho en comú amb la resta de la classe.

5. Louise Amoore destaca la importància de la transparència i l'ètica en l'ús dels algorismes. Per què creus que és important que els ciutadans comprenguin com funcionen els algorismes i quines dades s'utilitzen per entrenar-los? Com podem assegurar-nos que els algorismes s'utilitzin de manera justa i equitativa?

Entrevista a Geoffrey Hinton

Entrevista de Manuel G. Pascual publicada al diari *El Pais* el 7/5/2023.

—

Geoffrey Hinton anunció el lunes que ha renunciado a su puesto como vicepresidente de ingeniería de Google. Quiere dedicarse a alertar sobre el reverso tenebroso de la inteligencia artificial (IA), según dijo en una entrevista concedida a *The New York Times*. La suya no es una baja cualquiera: nacido en Wimbledon hace 75 años, este británico asentado en Canadá es una leyenda en la disciplina. Su trabajo ha sido decisivo para alumbrar algunas técnicas que han hecho posible ChatGPT, los traductores automáticos o los sistemas de visión de los vehículos autónomos. Hinton, que fue galardonado en 2017 con el premio Fronteras del Conocimiento que concede la Fundación BBVA, cree ahora que la tecnología que tanto ha ayudado a desarrollar puede llevarnos al fin de la civilización en cuestión de años.

La obsesión de este científico siempre fue estudiar cómo funciona el cerebro para tratar de replicar esos mecanismos en los ordenadores. En 1972 acuñó el concepto de red neuronal. La idea de fondo es aplicar matemáticas al análisis de datos para que el sistema sea capaz de desarrollar habilidades. Su propuesta no convenció en la época; hoy, las redes neuronales son la punta de lanza de la investigación en IA. El gran momento de Hinton llegó en 2012, cuando demostró el verdadero potencial de su línea de investigación con una red neuronal que podía analizar miles de fotografías y aprender por sí sola a distinguir ciertos objetos, como flores, coches o perros. También entrenó un sistema para que fuera capaz de predecir las siguientes letras de una frase inacabada, un antecesor de los actuales grandes modelos lingüísticos como el de ChatGPT.

Su trabajo le valió el Premio Turing, considerado el Nobel de la informática, que recibió en 2018 junto a otros investigadores como

2 https://elpais.com/tecnologia/2023-05-07/geoffrey-hinton-si-hay-alguna-forma-de-controlar-la-inteligencia-artificial-debemos-descubrirla-antes-de-que-sea-tarde.html

Yann LeCun, exalumno suyo, o Yoshua Bengio. También tiene en sus vitrinas el Premio Princesa de Asturias de Investigación Científica. Hinton atiende a EL PAÍS por videoconferencia desde su casa de Londres, adonde se ha trasladado tras dejar Google.

Geoffrey Hinton, en la sede de Google en Mountain View, California. Foto: Noah Berger

Pregunta. ¿Cuáles son los peligros de la IA a los que nos enfrentamos?
Respuesta. Hay muchos. La generación de noticias falsas ya está causando grandes divisiones en la sociedad. La eliminación de ciertos tipos de trabajo tendrá un impacto en el empleo. Aumentará la disparidad de riqueza entre los ricos y los pobres. Esos son algunos de los peligros inminentes, aunque yo no me centro en esos, sino en otro de carácter existencial. Hace poco me di cuenta de que el tipo de inteligencia digital que estamos desarrollando podría ser una forma de inteligencia mejor que la de los cerebros biológicos. Siempre pensé que la IA o el aprendizaje profundo intentaban imitar el cerebro, aunque no podían igualarlo: el objetivo era ir mejorando para que las máquinas se parecieran más y más a nosotros. He cambiado de postura en los últimos meses. Creo que podemos desarrollar algo que es mucho más eficiente que el cerebro porque es digital.

P. ¿Por qué lo cree?
R. El argumento es el siguiente. Con un sistema digital, podrías tener muchísimas copias de exactamente el mismo modelo del mundo.

Estas copias pueden funcionar en distintos hardwares. De este modo, diferentes copias podrían analizar datos diferentes. Y todas estas copias pueden saber al instante lo que las demás han aprendido. Lo hacen compartiendo sus parámetros. No podemos hacer eso con el cerebro. Nuestras mentes han aprendido a utilizar todas sus propiedades de forma individual. Si te diera un mapa detallado de las conexiones neuronales de mi cerebro, no te serviría de nada. Pero en los sistemas digitales, el modelo es idéntico. Todos usan el mismo conjunto de conexiones. Así, cuando uno aprende cualquier cosa, puede comunicárselo a los demás. Y es por eso que ChatGPT puede saber miles de veces más que cualquier persona: porque puede ver miles de veces más datos que nadie. Eso es lo que me asusta. Tal vez esta forma de inteligencia sea mejor que la nuestra.

P. Usted lleva décadas trabajando en esta disciplina. ¿Cómo ha llegado ahora a esta conclusión?
R. Ha sido al tratar de averiguar cómo un cerebro podría implementar los mismos procedimientos de aprendizaje que se utilizan en inteligencias digitales como las que están detrás de ChatGPT-4. Por lo que sabemos hasta ahora sobre el funcionamiento del cerebro humano, probablemente nuestro proceso de aprendizaje es menos eficiente que el de los ordenadores.

P. ¿Puede la IA ser realmente inteligente si no entiende lo que significan las palabras o sin tener intuición?
R. El aprendizaje profundo, si lo comparas con la IA simbólica [la corriente dominante en la disciplina hasta la irrupción de las redes neuronales, que trataba de que la máquina aprendiese palabras y números], es un modelo de intuición. Si tomas la lógica simbólica como referencia, si crees que así es como funciona el razonamiento, no puedes responder a la pregunta que te voy a hacer. Pero si tienes un modelo informático de intuición, la respuesta es obvia. Así que esta es la pregunta: sabes que hay gatos machos y hembras, y sabes que hay perros machos y hembras. Pero supongamos que te digo que tienes que elegir entre dos posibilidades, ambas ridículas: todos los gatos son machos y los perros son hembras, o todos los gatos son hembras y todos los perros son machos. En nuestra cultura,

tenemos bastante claro que tiene más sentido que los gatos sean hembras, porque son más pequeños, más listos y les rodean una serie de estereotipos, y que los perros sean machos, porque son más grandes, más estúpidos, más ruidosos, etcétera. Repito, no tiene ningún sentido, pero forzados a escoger, creo que la mayoría diría lo mismo. ¿Por qué? En nuestra mente representamos al gato y al perro, al hombre y a la mujer, con grandes patrones de actividad neuronal basándonos en lo que hemos aprendido. Y asociamos entre sí las representaciones que más se parecen. Ese es un razonamiento intuitivo, no lógico. Así es como funciona el aprendizaje profundo.

P. Usted sostiene que hasta ahora pensaba que la IA llegaría a superar a la inteligencia humana en unos 30 o 50 años. ¿Cuánto cree que queda ahora?
R. De cinco a 20 años
.
P. Eso está a la vuelta de la esquina.
R. No confío mucho en mi pronóstico porque creo que me equivoqué en el primero, pero está claro que todo se ha acelerado.

P. ¿Cree que la IA llegará a tener su propio propósito u objetivos?
R. Esa es una cuestión clave, quizás el mayor peligro que rodea a esta tecnología. Nuestros cerebros son el fruto de la evolución y tienen una serie de metas integradas, como no lastimar el cuerpo, de ahí la noción del daño; comer lo suficiente, de ahí el hambre; y hacer tantas copias de nosotros mismos como sea posible, de ahí el deseo sexual. Las inteligencias sintéticas, en cambio, no han evolucionado: las hemos construido. Por lo tanto, no necesariamente vienen con objetivos innatos. Así que la gran pregunta es, ¿podemos asegurarnos de que tengan metas que nos beneficien a nosotros? Este es el llamado problema del alineamiento. Y tenemos varias razones para preocuparnos mucho. La primera es que siempre habrá quienes quieran crear robots soldados. ¿O cree que Putin no los desarrollaría si pudiera? Eso lo puedes conseguir de forma más eficiente si le das a la máquina la capacidad de generar su propio conjunto de objetivos. En ese caso, si la máquina es inteligente, no tardará en darse cuenta

de que consigue mejor sus objetivos si se vuelve más poderosa. Deberíamos poner tanto esfuerzo en desarrollar esta tecnología como en asegurarnos de que sea segura

P. ¿Qué debemos hacer ahora?
R. Hay que llamar la atención de la gente sobre este problema existencial que supone la IA. Ojalá tuviera una solución, como en el caso de la emergencia climática: hay que dejar de quemar carbono, aunque haya muchos intereses que lo impidan. No conozco ningún problema equivalente al de la IA. Así que lo mejor que se me ocurre en este momento es que deberíamos poner tanto esfuerzo en desarrollar esta tecnología como en asegurarnos de que sea segura. Y eso no está sucediendo en la actualidad. ¿Cómo se logra eso en un sistema capitalista? Eso no lo sé.

P. ¿Cree que parte del problema reside en el hecho de que el desarrollo de la IA lo están llevando a cabo empresas privadas?
R. Así ha sido durante los últimos años. Google desarrolló internamente chatbots como LaMDA, que eran muy buenos, y deliberadamente decidió no abrirlos al público porque estaban preocupados por sus consecuencias. Y así fue mientras Google lideraba esta tecnología. Cuando Microsoft decidió poner un chatbot inteligente en su buscador Bing, Google tuvo que responder porque operan un sistema competitivo. Google se comportó de forma responsable, y no quiero que la gente piense que me fui para criticar a la compañía. Dejé Google para poder advertir sobre los peligros sin tener que pensar en el impacto que pueda causar en su negocio.

P. ¿Ha hablado de su decisión con otros colegas? ¿Tienen las mismas preocupaciones que usted?
R. Hemos entrado en un territorio completamente desconocido. Somos capaces de construir máquinas más fuertes que nosotros, pero aun así tenemos el control. ¿Qué pasa si desarrollamos máquinas más inteligentes que nosotros? No tenemos experiencia en tratar estas cosas. Hay gente a la que respeto, como mi colega Yann LeCun, que cree que lo que digo no tiene sentido. Sospecho que realmente tenemos que pensar mucho en esto. Y no basta con decir que no

vamos a preocuparnos. Muchas de las personas más inteligentes que conozco están seriamente preocupadas. Es lo que me ha convencido a dar un paso adelante y usar mi reputación para que la gente se dé cuenta de que se trata de un problema muy grave. No sirve de nada esperar a que la IA sea más lista que nosotros, debemos controlarla a medida que se desarrolla.

P. Usted no firmó la carta suscrita por más de un millar de expertos en IA que solicitaba una moratoria de seis meses en la investigación. ¿Por qué?
R. Creo que ese enfoque es completamente ingenuo. No hay manera de que eso suceda. Aun salvando la competencia entre las grandes empresas, está la de los países. Si EE UU decidiera dejar de desarrollar IA, ¿realmente cree que China se detendría? La idea de detener la investigación llama la atención de la gente sobre el problema, pero no va a suceder. Con las armas nucleares, dado que la gente se dio cuenta de que todos perderíamos si había una guerra nuclear, fue posible conseguir tratados. Con la IA será mucho más complicado porque es muy difícil comprobar si la gente está trabajando en ello.

P. ¿Qué propone, entonces?
R. Lo mejor que puedo recomendar es que muchas personas muy inteligentes traten de averiguar cómo contener los peligros de estas cosas. La IA es una tecnología fantástica, está provocando grandes avances en la medicina, en el desarrollo de nuevos materiales, en la previsión de terremotos o inundaciones... Necesitamos mucho trabajo para entender cómo contener la IA. No sirve de nada esperar a que la IA sea más lista que nosotros, debemos controlarla a medida que se desarrolla. También tenemos que comprender cómo contenerla, cómo evitar sus malas consecuencias. Por ejemplo, creo que todos los gobiernos deberían insistir en que todas las imágenes falsas lleven un distintivo.

P. ¿Es optimista sobre el futuro que nos aguarda?
R. Tiendo a ser una persona bastante optimista. Hay posibilidades de que no tengamos forma de evitar un mal final. Pero está claro

que también tenemos la oportunidad de prepararnos para este reto. Necesitamos mucha gente creativa e inteligente. Si hay alguna forma de mantener la IA bajo control, necesitamos descubrirla antes de que sea demasiado inteligente.

P. ¿Confía en que los gobiernos encuentren la forma de regular esta tecnología?
R. En Estados Unidos, el sistema político es incapaz de tomar una decisión tan simple como no dar fusiles de asalto a los adolescentes. Eso no aporta mucha confianza sobre cómo van a manejar un problema mucho más complicado como este.

P. El verano pasado, el ingeniero de Google Blake Lemoine se hizo famoso en todo el mundo al decir que el chatbot en el que trabajaba, LaMDA, había cobrado conciencia. ¿Fue ese caso premonitorio de lo que se nos venía encima?
R. Creo que lo que pasó encierra dos debates distintos. Primero, ¿se volverán las máquinas más inteligentes que nosotros para poder tomar el mando? Y segundo, ¿son conscientes o sensibles, o lo que sea que quiera decir eso? El debate más importante es el primero; en el segundo intervienen las creencias personales, y a mí eso no me interesa. En cualquier caso, me sorprende que haya muchísimas personas que están muy seguras de que las máquinas no son conscientes, pero que al mismo tiempo no sepan definir qué significa que algo o alguien sea consciente. Me parece una posición estúpida.

—

Preguntes:

1. Segons Geoffrey Hinton, quins són alguns dels perills actuals de la intel·ligència artificial?

2. Per què creu Hinton que la intel·ligència artificial desenvolupada podria arribar a ser millor que els cervells biològics?

3. Explica el concepte “aprenentatge profund” i en què es diferencia de la “intel·ligència artificial simbòlica”.

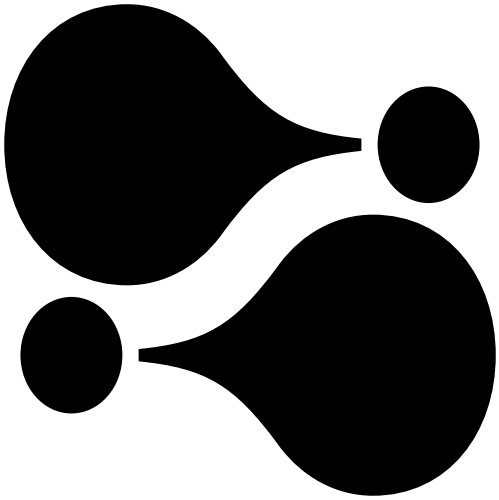
4. Què és el “problema d’alineació”, segons Hinton? Per què diu que suposa una “amenaça existencial”?

5. Per grups, feu un debat al voltant d’aquestes qüestions:

   - Quins són els aspectes positius i els negatius de la intel·ligència artificial?

   - Creus que els beneficis que té compensen els riscos que comporta?

   - S’hauria de regular el desenvolupament de la intel·ligència artificial? Si és que sí, com s’hauria de fer?



CCCB Centre de Cultura Contemporània de Barcelona