INTEL·LIGÈNCIA ARTIFICIAL: DE LA MÀGIA A LA REALITAT

Ramon López de Mántaras i Badia
Membre de l'Institut d'Estudis Catalans, de l'Institut d'Investigació en Intel·ligència Artificial i del Consell Superior d'Investigacions Científiques (Campus UAB, Bellaterra). mantaras@iiia.csic.es

Resum: Tot i els èxits indiscutibles de la intel·ligència artificial (IA) al llarg dels darrers anys, el fet és que els sistemes d'IA encara tenen una «intel·ligència» molt limitada, ja que són «intel·ligències» específiques, contràriament a la intel·ligència humana, que és general. L'excessiu antropocentrisme és el motiu principal que explica que la societat tingui una percepció errònia de l'estat real de la intel·ligència artificial. En aquest article expliquem quin és l'estat de la qüestió de la IA i també argumentem què cal fer per progressar cap a una IA més general i, per tant, més comparable amb la humana. Finalment, fem unes reflexions sobre el fet que, per més sofisticades que arribin a ser aquestes intel·ligències artificials, seran intel·ligències diferents de les nostres i, per tant, alienes a les necessitats i als valors humans. Això ens hauria de fer pensar sobre possibles limitacions ètiques al desenvolupament de la intel·ligència artificial.

Paraules clau: intel·ligència artificial, intel·ligència artificial general, ètica de la intel·ligència artificial.

ARTIFICIAL INTELLIGENCE: FROM MAGIC TO REALITY

Abstract: Despite the unquestionable achievements of artificial intelligence (AI) over the last few years, AI systems still have in fact a very limited "intelligence" as they are specific "intelligences" in contrast to the human intelligence, which is general. Excessive anthropocentrism is the main reason why society has a misperception of the state of the art of artificial intelligence. In this paper we explain what the real state of AI is and also argue what needs to be done to progress towards a more general AI which would, therefore, be more comparable to human intelligence. Lastly, we reflect on the fact that no matter how sophisticated they become, artificial intelligences will be different from ours and consequently alien to human values and needs. This should make us reflect on the possible ethical limitations to the development of artificial intelligence.

Keywords: artificial intelligence, general artificial intelligence, ethics of artificial intelligence.

Introducció

Imaginem que tenim una màquina per viatjar en el temps i que transportem Isaac Newton des de finals del segle XVII fins a l'actualitat, i que el situem en un lloc que li resulta familiar, com ara la capella del Trinity College, a la Universitat de Cambridge. Un cop allà, imaginem que li mostrem un mòbil d'última generació i el connectem. Sens dubte, ell, que va demostrar que la llum blanca es descompon en colors en incidir un raig de sol en un prisma, se sorprendria que un objecte tan petit produeixi colors tan vius a la foscor de la capella. Després fem que al mòbil soni una música que segurament reconeixeria, per exemple, una òpera de Händel. A continuació, li mostrem a la pantalla del dispositiu la seva obra *Principia mathematica* i li ensenyem com utilitzar dos dits per ampliar-ne el text. Suposem també que, tot seguit, li expliquem com fer fotos, gravar vídeos i so, fer càlculs aritmètics amb gran velocitat i precisió, comptar els passos que caminem, guiar-nos cap al nostre destí i, per descomptat, parlar amb algú a milers de quilòmetres. Newton seria capaç de donar una mínima explicació de com funciona un dispositiu tan meravellós? Tot i ser una de les ments més brillants de la història, que va inventar el càlcul infinitesimal i integral, va explicar tant l'òptica com la gravetat i va formular les lleis del moviment dels cossos que van revolucionar la física, seria incapaç de donar-hi una explicació mínimament coherent. No podria distingir aquest dispositiu de la màgia. Ens podríem també preguntar què més s'imaginaria Newton que aquest dispositiu pot fer? Creuria que pot funcionar indefinidament? —recordem que va viure en una època cent anys anterior a Alessandro Volta, l'inventor de la pila elèctrica—, creuria que pot transformar plom en or? —recordem que la química de la seva època era l'alquímia. Possiblement sí, ja que tendim a no veure els límits a allò que ens sembla màgic. Aquest és un dels problemes que tenim a l'hora de comprendre tecnologies molt avançades. Ja ho va dir Arthur Clarke als anys seixanta: «Qualsevol tecnologia prou sofisticada no es pot distin-

gir de la màgia». Amb la intel·ligència artificial (IA) passa el mateix. Sembla que no hi hagi límits en el seu potencial però, en realitat, la IA roman encallada des de fa més de cinquanta anys en una de les qüestions més fonamentals: com dotar les màquines de coneixements de sentit comú? És una qüestió crucial si volem assolir intel·ligències artificials de tipus general indistingibles de la intel·ligència humana. Fins avui els investigadors en IA no hem vist cap indici que ens porti a poder afirmar que aquest problema pugui ser resolt ni a curt ni a mitjà termini. De fet, l'Agència d'Investigacions de Projectes Avançats de Defensa (DARPA, Defense Advanced Research Projects Agency), la institució que més inverteix en programes de recerca als Estats Units d'Amèrica, a finals del 2018 va anunciar (DARPA, 2018) que finançaria amb dos mil milions de dòlars un programa de recerca sobre com dotar les màquines de coneixements de sentit comú. L'absència de coneixements de sentit comú impossibilita que un sistema d'IA pugui comprendre el llenguatge, pugui entendre allò que percep mitjançant els seus sensors, pugui gestionar bé situacions imprevistes i pugui aprendre a partir de l'experiència. Resoldre el problema de l'adquisició de coneixements de sentit comú seria un gran avenç de la intel·ligència artificial, ja que obriria la porta al desenvolupament d'intel·ligències artificials de tipus general i se superarien així les limitacions actuals de la IA específica, és a dir, capaç de dur a terme només una sola tasca.

La realitat de la IA

Quina és, doncs, la situació real de la IA? La realitat és que el que tenim són «intel·ligències» summament específiques en el sentit que cadascuna sap fer bé una única tasca. Focalitzem-nos en una tècnica d'IA, coneguda com a *aprenentatge profund*, que bàsicament consisteix en xarxes neuronals artificials multinivell (Bengio, 2009) que han aconseguit espectaculars resultats recentment, com, per exemple, un programari anomenat AlphaZero (Silver *et al.*, 2018), que, jugant contra si mateix milions de partides durant hores, va aprendre a jugar a *go* a uns nivells que no s'havien aconseguit mai abans i va superar amb escreix els millors jugadors humans. Doncs bé, aquests sistemes d'aprenentatge profund són summament limitats, ja que només són capaços d'aprendre a detectar patrons analitzant enormes quantitats de dades. No és exagerat afirmar que, de fet, no aprenen realment res; almenys en el sentit humà del que entenem per aprendre. És a dir, en realitat, no saben res de nou després d'haver estat entrenats per adquirir una competència. N'és una prova el que es coneix com a *oblit catastròfic*, que significa que els sistemes d'aprenentatge profund obliden tot allò que s'ha après prèviament a partir de l'instant que se'ls ensenya alguna cosa nova. Per exemple, si després d'haver «après» a jugar a *go* entrenem un sistema d'aprenentatge profund a diferenciar entre gats i gossos, després de mostrar-li molts milers d'imatges, aprendrà perfectament a distingir-los, però serà incapaç de tornar a jugar a *go*. Caldria tornar a entrenar-lo perquè de nou «aprengués a jugar a *go*», la qual cosa provocaria que, a continuació, seria incapaç de distingir els gats dels gossos. En altres paraules, contràriament a nosaltres, els sistemes d'IA no aprenen de manera incremental ni poden relacionar allò novament après amb el que havien après anteriorment. A més, nosaltres no necessitem veure milers de gats i de gossos per distingir-los, amb uns quants en tenim prou. A mitjà termini és possible aconseguir desenvolupar sistemes intel·ligents més generalistes, és a dir, no limitats com ara a resoldre una única tasca, sinó capaços d'executar-ne amb excel·lència diverses alhora, i molt possiblement seran sistemes que combinaran components d'aprenentatge basat en l'anàlisi de dades amb components de raonament basats en coneixements representats mitjançant llenguatges de representació que tindran com a base la lògica matemàtica i les seves extensions.

Quin és, doncs, el motiu pel qual molts creuen que la IA està a punt d'igualar la intel·ligència humana i, a partir d'aquesta falsa premissa, fan prediccions sobre una possible singularitat tecnològica? (és a dir, el moment en què la superintel·ligència artificial farà que la intel·ligència humana sigui àmpliament superada). Al meu entendre, l'excessiu antropocentrisme és el principal motiu pel qual la societat té una percepció errònia de l'estat real de la intel·ligència artificial. Quan ens informen d'èxits espectaculars d'una IA específica en una competència molt complexa, encara que sigui molt concreta, tendim a generalitzar i li atribuïm la capacitat de fer pràcticament qualsevol cosa que fem els éssers humans i, fins i tot, de fer-la molt millor. En altres paraules, creiem que la IA pràcticament no té límits quan, de fet, és extremament limitada i, el que és molt important, gairebé no té res a veure amb la intel·ligència humana; en realitat, el que tenen els actuals sistemes d'IA no és «intel·ligència» sinó «habilitats sense comprensió» en el sentit que apunta Daniel Dennet al seu llibre *From bacteria to Bach and back* (Dennet, 2017). És a dir, són sistemes que poden arribar a ser molt hàbils duent a terme tasques específiques, com discriminar una sèrie d'elements en una imatge, però sense comprendre absolutament res sobre la naturalesa d'aquests elements ni de les propietats ni de les relacions entre ells a causa de l'absència de sentit comú. Per exemple, poden identificar una persona davant d'una paret, però no saben que les persones no poden travessar parets ni que les persones no poden ser a dos llocs alhora. Sense aquests coneixements no és possible una comprensió profunda del llenguatge ni una interpretació profunda del que capta un sistema de percepció visual, entre d'altres limitacions. De fet, com ja hem dit, el sentit comú és un requisit fonamental per assolir una IA similar a la humana quant a generalitat i profunditat. Els coneixements de sentit comú són fruit de les nostres vivències i experiències. Alguns exemples són: «l'aigua sempre flueix de dalt a baix», «per arrossegar un objecte lligat a una corda cal estirar la corda, no empènyer-la», «un got es pot guardar dins d'un armari, però no podem guardar un armari dins un got», etc. Hi ha milions de coneixe-

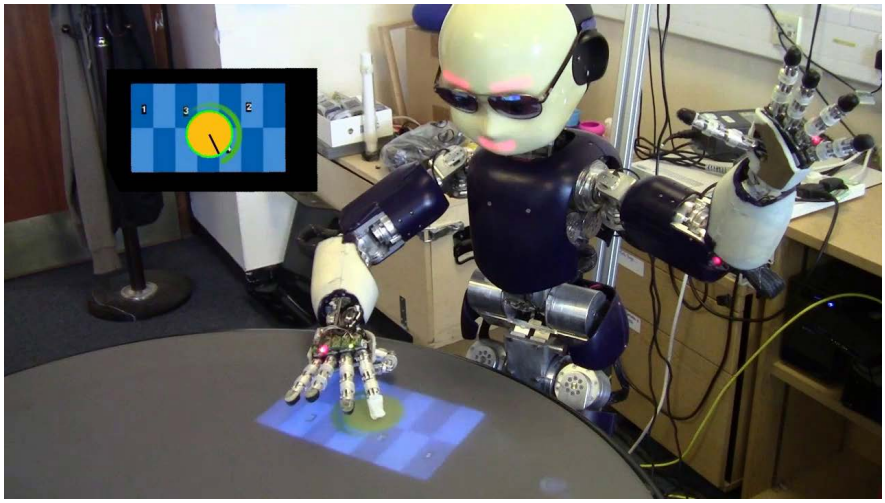

FIGURA 1. Robot hominoide que aprèn a relacionar la posició del dit en un teclat virtual amb la nota musical que sona.
FONT: Institut d'Investigació en Intel·ligència Artificial (IIIA).

ments de sentit comú que les persones fem servir fàcilment i que ens permeten entendre el món en què vivim. Una possible línia de recerca que podria donar resultats interessants en adquisició de coneixements de sentit comú és allò que es coneix en el camp de la robòtica com a *robòtica del desenvolupament*, que està basada en els estudis de Jean Piaget sobre el desenvolupament cognitiu, és a dir, el procés pel qual els infants aprenen interactuant amb l'entorn. Una altra línia de treball molt interessant és la que té com a objectiu la modelització matemàtica i l'aprenentatge de relacions causa-efecte, és a dir, l'aprenentatge de models causals, i per tant asimètrics, del món (figura 1). Els sistemes actuals basats en aprenentatge profund simplement poden aprendre funcions matemàtiques simètriques, no poden aprendre relacions asimètriques i, per tant, no són capaços de diferenciar entre causes i efectes, com ara que la sortida del sol és la causa del cant del gall i no el contrari.

Futur: cap a intel·ligències artificials realment intel·ligents

Les capacitats més complicades d'assolir són aquelles que requereixen interaccionar amb entorns no restringits ni prèviament preparats. Dissenyar sistemes que tinguin aquestes capacitats requereix integrar desenvolupaments a moltes àrees de la IA. En particular, necessitem llenguatges de representació de coneixements que codifiquin informació sobre molts tipus diferents d'objectes, situacions, accions, etc., així com de les seves propietats i les relacions entre ells, en particular, relacions causa-efecte. També necessitem nous algorismes que, a partir d'aquestes representacions, puguin, de manera robusta i eficient, resoldre problemes i respondre a preguntes sobre pràcticament qualsevol tema. Finalment, atès que necessitaran adquirir un nombre pràcticament il·limitat de coneixements, aquests sistemes hauran de ser capaços d'aprendre de manera contínua al llarg de tota la seva existència. En definitiva, és imprescindible dissenyar sistemes que integrin components de la intel·ligència tals com percepció, representació, raonament, acció i aprenentatge. Aquest és un problema molt important en IA, ja que encara no sabem com integrar tots aquests components de la intel·ligència. Necessitem arquitectures que permetin integrar eficientment aquests components de forma adequada. Aquestes arquitectures s'anomenen *arquitectures cognitives*.

Entre les activitats futures, creiem que els temes de recerca més importants passaran per sistemes híbrids que combinin els avantatges que tenen els sistemes capaços de raonar sobre la base de coneixements i ús de la memòria i els avantatges de la IA basada en l'anàlisi de quantitats massives de dades, és a dir, en allò que es coneix com a *aprenentatge profund* (Bengio, 2009). A més del problema de l'oblit catastròfic esmentat anteriorment, una altra limitació important d'aquests sistemes és que són «caixes negres» sense capacitat explicativa, per això un objectiu interessant de recerca serà com dotar de capacitat explicativa els sistemes d'aprenentatge profund incorporant mòduls que permetin explicar com s'ha arribat als resultats i les conclusions proposats, ja que la capacitat d'explicació és una característica irrenunciable en qualsevol sistema intel·ligent. També cal desenvolupar nous algorismes d'aprenentatge que no requereixin enormes quantitats de dades per ser entrenats, així com maquinari molt més eficient en consum energètic per implementar-los, ja que el consum d'energia podria acabar sent una de les barreres principals al desenvolupament de la IA. En comparació, el cervell és diversos ordres de magnitud més eficient que el maquinari actual necessari per implementar els algorismes d'IA més sofisticats.

Altres tècniques més clàssiques d'IA que continuaran sent objecte de recerca extensiva són els sistemes multiagent (és a dir, sistemes que es componen d'un conjunt d'agents intel·ligents amb experteses complementàries que col·laboren per resoldre problemes que requereixen col·laboració entre experts per ser resolts), la planificació d'accions, el raonament basat en l'experiència, la visió artificial, la comunicació multimodal persona-màquina, la robòtica humanoide i, especialment, les noves tendències en robòti-

ca del desenvolupament, que, com hem esmentat, poden ser la clau per dotar les màquines de sentit comú i, en particular, aprendre la relació entre les accions d'aquestes mateixes màquines i els efectes que produeixen a l'entorn. També veurem progressos significatius gràcies a les aproximacions biomimètiques per reproduir en màquines el comportament d'animals. No es tracta únicament de reproduir el comportament d'un animal, sinó de comprendre com funciona el cervell que produeix aquest comportament.

Pel que fa a les aplicacions, algunes de les més importants continuaran sent aquelles relacionades amb la World Wide Web, els videojocs, els assistents personals i els robots autònoms (en particular, vehicles autònoms, robots socials, robots per a l'exploració de planetes, etc.). Les aplicacions al medi ambient i l'estalvi energètic també seran importants, així com les de l'economia i la sociologia.

Finalment, les aplicacions de la intel·ligència artificial a l'art (arts visuals, música, dansa, narrativa) canviaran de manera important la naturalesa del procés creatiu. Els ordinadors ja no són només eines d'ajuda a la creació, sinó que comencen a ser agents creatius. Això ha donat lloc a una nova i molt prometedora àrea d'aplicació de la intel·ligència artificial anomenada *creativitat computacional* que ja ha produït resultats molt interessants (Colton, López de Mántaras i Stock, 2009; Colton *et al.*, 2015; López de Mántaras, 2016) en escacs, música, arts plàstiques i narrativa, entre altres activitats creatives.

Algunes reflexions finals

La intel·ligència humana és el referent principal de cara a assolir l'objectiu últim de la IA, és a dir, la IA general comparable o fins i tot superior a la intel·ligència humana, però al meu entendre, per més sofisticada que arribi a ser la IA, sempre serà diferent de la humana, ja que el desenvolupament mental que requereix tota intel·ligència complexa depèn de les interaccions amb l'entorn, i aquestes interaccions depenen alhora del cos, en particular del sistema perceptiu i del sistema motor. Això, juntament amb el fet que les màquines molt probablement no seguiran processos de socialització i culturització, incideix encara més en el fet que, per més sofisticades que arribin a ser, seran intel·ligències diferents de les nostres. El fet de ser intel·ligències alienes a la intel·ligència humana i, per tant, alienes a les necessitats i els valors humans ens hauria de fer reflexionar sobre possibles limitacions ètiques al desenvolupament de la intel·ligència artificial. En particular, estem d'acord amb l'afirmació de Weizenbaum (1976) que cap màquina no hauria de prendre mai decisions de forma completament autònoma o donar consells que requereixin, entre altres coses, la saviesa, producte d'experiències humanes, així com tenir en compte valors humans. És a dir, el perill de la IA no és la singularitat tecnològica deguda a l'existència d'unes futures hipotètiques superintel·ligències artificials; els veritables problemes ja són aquí i tenen a veure amb la privadesa i la vigilància i el control massiu de la ciutadania, l'autonomia dels sistemes (en particular les armes autònomes), la confiança excessiva sobre les capacitats de la IA, el biaix dels algorismes d'aprenentatge i la impossibilitat de retre comptes per justificar les seves decisions en un llenguatge comprensible per a les persones.

Considerem, per començar, el problema de la privadesa: actualment els algorismes en què es basen els motors de cerca a Internet, els sistemes de recomanació i els assistents personals dels nostres telèfons mòbils, coneixen força bé el que fem, les nostres preferències i els nostres gustos i, fins i tot, poden arribar a inferir allò que pensem i com ens sentim. L'accés a quantitats massives d'informació que generem voluntàriament és fonamental perquè això sigui possible, ja que mitjançant l'anàlisi d'aquestes dades provinents de fonts diverses és possible trobar relacions i patrons que serien impossibles de detectar sense les tècniques d'IA. Tot això resulta en una pèrdua alarmant de privadesa. Per intentar evitar-ho hauríem de tenir dret a posseir una còpia de totes les dades personals que generem, controlar-ne l'ús i decidir a qui permetem accedir-hi i sota quines condicions en lloc que estiguin en mans de grans corporacions.

Continuem amb el problema de l'autonomia: la IA està basada en programació complexa, i, per tant, necessàriament cometrà errors. Però fins i tot suposant que fos possible desenvolupar un programari completament fiable, hi ha dilemes ètics que els desenvolupadors de programari han de tenir en compte a l'hora de dissenyar sistemes autònoms. Per exemple, un vehicle autònom podria decidir atropellar un vianant per evitar una col·lisió que podria causar danys als ocupants. Un altre exemple clar són les armes autònomes. Els tres principis bàsics que regeixen els conflictes armats: discriminació (la necessitat de discernir entre combatents i civils o entre un combatent rendint-se i un en disposició d'atacar), proporcionalitat (fins a quin punt són acceptables els danys col·laterals) i precaució (minimització del nombre de víctimes) són extraordinàriament difícils d'avaluar i, per tant, són gairebé impossibles de complir pels sistemes d'IA que controlen les armes autònomes. Però, fins i tot en el cas que, a molt llarg termini, les màquines tinguessin aquestes capacitats, al meu parer seria indigne delegar en una màquina la decisió de matar. En general, com més autonomia donem als sistemes d'IA, més responsabilitat hauríem d'exigir als dissenyadors i programadors d'aquests sistemes, de manera que compleixin principis legals i ètics. És a dir, el veritable problema no és el monstre de Frankenstein sinó el doctor Frankenstein.

Pel que fa a l'impacte al mercat laboral, tot i ser cert que serà important, possiblement no ho sigui tant com alguns preveuen. Sens dubte, l'innegable entusiasme actual per la IA ens pot fer creure que la intel·ligència humana és substituïble, i això ha portat algunes organitzacions a acomiadar empleats i reemplaçar-los per sistemes d'IA. Això és un error greu, ja que, de fet, tots els sistemes d'IA depenen críticament de la intel·ligència humana. Els sistemes

basats en el coneixement es fonamenten en el coneixement i la comprensió de l'experiència humana, i els sistemes d'IA que es basen en dades depenen críticament de dades sobre la conducta humana. D'aquí es desprèn que cal continuar ensenyant, desenvolupant i exercint la capacitat humana. D'altra banda, en la gran majoria de casos, la capacitat humana encara supera amb escreix la intel·ligència artificial, especialment quan el sistema d'IA s'ha d'enfrontar a situacions que no han aparegut en els conjunts de dades amb què s'han entrenat els sistemes d'IA. A més, sovint moltes aplicacions es beneficien de la sinergia entre l'ésser humà i la intel·ligència artificial, és a dir, la unió persona-màquina està produint resultats superiors a qualsevol dels dos per separat, ja que per més dades que pugui analitzar una màquina, sempre caldrà el judici humà; un dels motius és que les màquines no poden distingir entre correlació i causalitat. Aquest fenomen es dona en àmbits com el diagnòstic mèdic i la presa de decisions en general, incloses les decisions empresarials. Els treballadors necessitaran rebre una formació continuada que els permeti adaptar-se a noves formes de treball que requeriran més creativitat, col·laboració entre ells i amb les màquines, i iniciativa per a llocs de treball canviants organitzats per a tasques concretes amb una elevada mobilitat tant geogràfica com funcional. Les empreses, per part seva, hauran d'invertir molt més en IA i, en particular, en la formació contínua dels seus empleats, incloent-hi els executius. Encara poques empreses han incorporat la IA a la seva cadena de valor; un dels principals factors limitatius és l'escassetat de persones amb una formació adequada en IA.

Un altre problema són els biaixos dels algorismes. Cap sistema d'IA no té intencionalitat, però les decisions que prenen estan basades en dades d'entrenament que sovint estan esbiaixades, de manera que les decisions que prenen estan també esbiaixades. La IA no només reprodueix els biaixos humans, sinó que els amplifica. Per exemple, un sistema de preselecció de candidats a un lloc de treball de nivell directiu va ser entrenat amb dades històriques que reflectien estadísticament que els executius més reeixits eren homes blancs, per la qual cosa el sistema discriminava candidats dones i afroamericans. El problema és que l'algorisme no tenia en compte la minoritària presència de dones i persones no blanques a les dades d'entrenament de l'algorisme. Un altre exemple és un sistema d'anàlisi d'imatges que, després de ser entrenat amb milers d'imatges quotidianes, va associar les imatges de dones amb imatges de cuines, però imatges d'homes amb activitats esportives. Un altre exemple, potser més preocupant, és el programari Compass, usat per jutges als Estats Units d'Amèrica per avaluar la probabilitat de reincidència, que atorgava una probabilitat doblement superior de reincidir a ciutadans afroamericans que a ciutadans blancs (Corbett-Davis *et al.*, 2016). Aquest biaix va ser detectat i denunciat i va haver de ser corregit, però ha influït en moltes decisions de jutges abans de la correcció. És necessari establir metodologies de verificació i validació adequades dels algorismes d'IA per tal que siguin utilitzades per autoritats certificadores, com també els processos de certificació de la seguretat dels aliments que consumim o els medicaments que prenem.

El darrer problema que volia esmentar és la rendició de comptes dels algorismes. Quan un sistema d'IA pren decisions, les persones afectades han de poder rebre una explicació de per què es pren la decisió en un llenguatge comprensible i han de ser capaces de qüestionar-la amb arguments raonats. Això és especialment important en camps com ara decisions sobre préstecs, sentències legals (per exemple, en la concessió d'una llibertat condicional), assegurances, impostos, etc. Molts sistemes d'intel·ligència artificial, especialment els que es basen en enfocaments a partir de dades, actualment no poden proporcionar aquest tipus d'explicació. Les seves decisions es deriven d'un conjunt ampli de paràmetres obtinguts estadísticament. Som als inicis de les investigacions sobre tècniques per comprendre el funcionament d'aquests sistemes, i és probable que, de nou, calgui una combinació d'IA basada en el coneixement i d'IA basada en les dades. La rendició de comptes és clarament una condició prèvia a qualsevol desplegament racional d'aplicacions de la IA.

Conclusions

La IA i els seus algorismes no són neutrals, sinó el reflex de les intencions i els biaixos de l'equip de programadors i d'entitats implicats en la seva implementació, amb l'afegit que no només els reflecteixen sinó que els amplifiquen. Tots aquests problemes relacionats amb l'impacte de la IA fan que molts experts assenyalem la necessitat de regular-ne el desenvolupament.

Però, a més de regular, és imprescindible educar la ciutadania sobre els beneficis i riscos de les tecnologies intel·ligents (que no són els que veiem a les pel·lícules de ciència-ficció), dotant (els ciutadans) de les competències necessàries per controlar-les en lloc de ser controlats per elles. Necessitem futurs ciutadans molt més informats, amb més capacitat per avaluar els riscos tecnològics, amb molt més sentit crític i capaços de fer valer els seus drets. Aquest procés de formació ha de començar a les escoles i tenir continuació a la universitat. En particular, cal que l'estudiantat de ciència i enginyeria rebi una formació ètica que li permeti comprendre millor les implicacions socials de les tecnologies que desenvoluparà. Només si invertim en educació aconseguirem una societat que pugui aprofitar els avantatges de les tecnologies intel·ligents minimitzant-ne els riscos, i així la intel·ligència artificial servirà per fer un gran pas en el progrés de la humanitat.

Bibliografia

Bengio, Y. (2009). «Learning deep architectures for AI». *Foundations and Trends in Machine Learning*, vol. 2, núm. 1, p. 1-127.

Colton, S.; Halskov, J.; Ventura, D.; Gouldstone, I.; Cook, M.; Pérez-Ferrer, B. (2015). «The painting fool sees! New projects with the automated painter». A: *Proceedings of the Sixth International Conference on Computational Creativity* (*ICCC 2015*). Provo, Utah: Brigham Young University, p. 189-196.

Colton, S.; López de Mántaras, R.; Stock, O. (2009). «Computational creativity: Coming of age». *AI Magazine*, vol. 30, núm. 3, p. 11-14.

Corbett-Davies, S.; Pierson, E.; Feller, A.; Goel, S. (2016). «A computer program used for bail and sentencing decisions was labeled biased against blacks. It's actually not that clear». *The Washington Post* (17 octubre).

Defense Advanced Research Projects Agency (DARPA) (2018). «Teaching machines common sense reasoning» [en línia]. <https://www.darpa.mil/news-events/2018-10-11>.

Dennet, D. C. (2017). *From bacteria to Bach and back*. Londres: Penguin Books.

López de Mántaras, R. (2016). «Artificial intelligence and the arts: Toward computational creativity». A: *The next step: Exponential Life*. Madrid: BBVA Open Mind, p. 100-125.

Silver, D.; Hubert, T.; Schrittwieser, J.; Antonoglou, I.; Lai, M.; Guez, A.; Lanctot, M.; Sifre, L.; Kumaran, D.; Graepel, T.; Lillicrap, T.; Simonyan, K.; Hassabis, D. (2018). «A general reinforcement learning algorithm that masters chess, shogi, and go through self-play». *Science*, vol. 362, núm. 64199, p. 1140-1144.

Weizenbaum, J. (1976). *Computer power and human reasoning: From judgment to calculation*. San Francisco: W. H. Freeman and Co.