MONOGRÀFIC
MÈTODE SCIENCE STUDIES JOURNAL (2018). Universitat de València.
DOI: 10.7203/metode.9.11145
Article rebut: 12/12/2017, acceptat: 23/07/2018.

CAP A LA INTEL·LIGÈNCIA ARTIFICIAL

PROGRESSOS, REPTES I RISCOS

RAMON LÓPEZ DE MÁNTARAS

Aquest article conté algunes reflexions al voltant de la intel·ligència artificial (IA). En primer lloc es fa la distinció entre IA forta i feble i els conceptes relacionats d'IA general i específica. A continuació, es descriuen breument els principals models existents. També es discuteix la necessitat de poder dotar de coneixements de sentit comú a les màquines per avançar cap a l'objectiu de construir IA general. Després parlem de les tendències en IA basada en l'anàlisi de grans quantitats de dades que han permès assolir progressos espectaculars molt recentment. Per finalitzar, parlem d'altres temes que són i continuaran essent clau en IA i tanquem amb una breu reflexió sobre els riscos de la IA.

Paraules clau: intel·ligència artificial forta, intel·ligència artificial feble, coneixements de sentit comú, aprenentatge profund.

INTRODUCCIÓ

L'objectiu últim de la intel·ligència artificial (IA), aconseguir que una màquina tingui una intel·ligència de tipus general similar a la humana, és una de les fites més ambicioses que s'ha plantejat la ciència. És d'una dificultat comparable a altres grans objectius científics com ara explicar l'origen de la vida o l'origen de l'univers, o bé conèixer l'estructura de la matèria. Al llarg dels darrers segles, aquest afany per construir màquines intel·ligents ens ha conduït a inventar models o metàfores del cervell humà. Per exemple, al segle XVII, Descartes es va preguntar si un complex sistema mecànic compost d'engranatges, politges i tubs podria, en principi, emular el pensament. Dos segles després, la metàfora es va plasmar en els sistemes telefònics, ja que semblava que les seves connexions es podien assimilar a una xarxa neuronal. Actualment el model dominant és el model computacional basat en l'ordinador digital i per tant és el model al qual ens referirem en aquest article.

> «LA IA ÉS EL CAMP CIENTÍFIC DEDICAT A VERIFICAR SI UN ORDINADOR ÉS CAPAÇ O NO DE TENIR UNA CONDUCTA INTEL·LIGENT DE TIPUS GENERAL»

INTEL·LIGÈNCIA ARTIFICIAL FEBLE 'VERSUS' FORTA

Allen Newell i Herbert Simon van formular la hipòtesi segons la qual tot sistema de símbols físics posseeix els mitjans necessaris i suficients per dur a terme accions intel·ligents (Newell i Simon, 1976). D'altra banda, com que els éssers humans som capaços de mostrar conductes intel·ligents, d'acord amb la hipòtesi, nosaltres som també sistemes de símbols físics. Convé aclarir a què es refereixen Newell i Simon. Un sistema de símbols físics consisteix en un conjunt d'entitats anomenades símbols que, mitjançant relacions, poden ser combinats per formar estructures més grans –com els àtoms que es combinen formant molècules– i que poden ser transformats aplicant un conjunt de procediments. Aquests procediments poden crear nous símbols, crear i modificar relacions entre aquests, emmagatzemar-ne, comparar si dos són iguals o diferents, etcètera. Aquests símbols són físics en tant que tenen un substrat fisicoelectrònic (en el cas dels ordinadors) o fisicobiològic (en el cas dels éssers humans). Efectivament, en el cas dels ordinadors els símbols es realitzen mitjançant circuits electrònics digitals i en el cas dels éssers humans, mitjançant xarxes de neurones. En definitiva, d'acord amb la hipòtesi del sistema de símbols físics, la naturalesa del substrat (circuits electrònics o xarxes neuronals) no té importància sempre que aquest permeti processar símbols. No oblidem que es tracta d'una hipòtesi i per tant la seva validesa o re-

futació s'haurà de verificar d'acord amb el mètode científic. La intel·ligència artificial és precisament el camp científic dedicat a intentar verificar aquesta hipòtesi en el context dels ordinadors, és a dir, verificar si un ordinador convenientment programat és capaç o no de tenir una conducta intel·ligent de tipus general.

És important el matís que s'hauria de tractar d'intel·ligència de tipus general, i no una d'específica, ja que la intel·ligència dels éssers humans és de tipus general. Exhibir intel·ligència específica és una altra cosa ben diferent. Per exemple, els programes que juguen als escacs al nivell de gran mestre són incapaços de jugar a les dames. Es requereix un programa diferent perquè el mateix ordinador jugui a les dames; és a dir, aquest no pot aprofitar el fet que juga als escacs per adaptar-se i jugar també a les dames. En el cas dels éssers humans, qualsevol jugador d'escacs pot aprofitar els seus coneixements sobre aquest joc per a jugar a les dames perfectament. La intel·ligència artificial que únicament mostra comportament intel·ligent en un àmbit molt específic està relacionada amb el que es coneix com a «IA feble» en contraposició amb la «IA forta» a la qual, de fet, es referien Newell i Simon i altres pares fundadors de la IA.

Qui va introduir aquesta distinció entre IA feble i forta fou el filòsof John Searle en un article crític amb la intel·ligència artificial publicat el 1980 (Searle, 1980) que va provocar, i continua provocant, molta polèmica. La IA forta implicaria que un ordinador convenientment programat no simula una ment sinó que «és una ment» i per tant hauria de ser capaç de pensar igual que un ésser humà. Searle en el seu article intenta demostrar que la IA forta és impossible.

En aquest punt convé aclarir que no és el mateix IA general que IA forta. Hi ha, òbviament, una connexió però només en un sentit: és a dir, que tota IA forta serà necessàriament general però hi pot haver IA generals que no siguin fortes, això és, que simulin la capacitat d'exhibir intel·ligència general de la ment però sense ser ments.

La IA feble, d'altra banda, consistiria, segons Searle, a construir programes que realitzin tasques específiques. La capacitat dels ordinadors per a realitzar tasques específiques fins i tot millor que les persones ja s'ha demostrat àmpliament en certs dominis, com ara buscar solucions a fórmules lògiques amb moltes variables i altres aspectes relacionats amb la presa de decisions. També s'associa amb la IA feble el fet de formular i provar hipòtesis sobre aspectes relacionats amb la ment (per exemple, la capacitat de raonar deductivament, d'aprendre inductivament, etc.) mitjançant la construcció de programes que duen a terme aquestes funcions encara que sigui mitjançant processos completament diferents dels que duu a terme el cervell. Absolutament tots els avenços aconseguits fins ara en el camp de la IA són manifestacions de la IA feble i específica.

«TOTS ELS AVENÇOS ACONSEGUITS FINS ARA SÓN MANIFESTACIONS DE LA IA FEBLE I ESPECÍFICA»

ELS PRINCIPALS MODELS EN INTEL·LIGÈNCIA ARTIFICIAL

Fins molt recentment, el model dominant en IA ha estat el simbòlic. Aquest model té les arrels en la hipòtesi del sistema de símbols físics. Encara continua sent molt important i actualment es considera el model «clàssic» en IA. És un model descendent (*top-down*) que es basa en el raonament lògic i la cerca heurística com a pilars per a la resolució de problemes, sense que el sistema intel·ligent necessiti formar part d'un cos ni estar situat en un entorn

Universitat Carnegie Mellon

En la dècada dels setanta, Allen Newell i Herbert Simon van plantejar que tot sistema de símbols físics –ja siguen aquests fisicoelectrònics en el cas dels ordinadors o fisicobiològics en el dels éssers humans– posseeix els mitjans necessaris per a dur a terme accions intel·ligents. En la imatge, els professors Simon (a l'esquerra) i Newell (a la dreta), treballant en la programació d'escacs a finals dels anys cinquanta, a la Universitat Carnegie Mellon a Pittsburgh (EUA).

IBM

Els éssers humans posseeixen una intel·ligència de tipus general, mentre que els programes que juguen als escacs al nivell de gran mestre, com és el cas de l'ordinador Deep Blue que va aconseguir guanyar el campió Kasparov l'any 1997, tenen una intel·ligència de tipus específic. Això vol dir que són incapaços d'utilitzar els seus coneixements per a jugar, per exemple, a les dames. En la imatge, l'equip d'IBM que va desenvolupar l'ordinador Deep Blue, en una imatge de 1996.

real. És a dir, la IA simbòlica opera amb representacions abstractes del món real que es modelen mitjançant llenguatges de representació basats principalment en la lògica matemàtica i les seves extensions. Per aquest motiu, els primers sistemes intel·ligents resolien sobretot problemes que no requerien interactuar directament amb l'entorn, com ara demostrar senzills teoremes matemàtics o jugar a escacs. Això no vol dir que la IA simbòlica no es pugui usar per programar el mòdul de raonament d'un robot físic situat en un entorn real, però en els primers anys de la IA no hi havia llenguatges de representació del coneixement ni de programació que permetessin fer-ho de forma eficient. Actualment, la IA simbòlica se segueix usant per demostrar teoremes o jugar a escacs, però també per a aplicacions que requereixen percebre l'entorn i actuar sobre aquest, com ara l'aprenentatge i la presa de decisions en robots autònoms.

«LA CAPACITAT DELS ORDINADORS PER A REALITZAR TASQUES ESPECÍFIQUES FINS I TOT MILLOR QUE LES PERSONES JA S'HA DEMOSTRAT ÀMPLIAMENT»

Simultàniament amb la IA simbòlica, també va començar a desenvolupar-se una IA bioinspirada anomenada connexionista. Contràriament a la IA simbòlica, es tracta d'una modelització ascendent (*bottom-up*), ja que es basa en la hipòtesi que la intel·ligència emergeix a partir de l'activitat distribuïda d'un gran nombre d'unitats interconnectades que processen informació paral·lelament. En la IA connexionista aquestes unitats són models molt aproximats de l'activitat elèctrica de les neurones biològiques. McCulloch i Pitts (1943) van proposar un model simplificat de neurona d'acord amb la idea que aquesta és essencialment una unitat lògica. Aquest model és una abstracció matemàtica amb entrades i sortides, que es correspondrien, respectivament, amb les dendrites i els axons. El valor de la sortida es calcula en funció del resultat d'una suma ponderada de les entrades, de manera que si aquesta suma supera un llindar preestablert llavors la sortida és un 1; en cas contrari, la sortida és 0. Connectant la sortida de cada neurona amb les entrades d'altres neurones es forma una xarxa neuronal artificial. D'acord amb el que ja se sabia llavors sobre el reforçament de les sinapsis entre neurones biològiques, es va veure que aquestes xarxes neuronals artificials es poden entrenar perquè aprenguin funcions que relacionin les entrades amb les sortides mitjançant l'ajust dels pesos que serveixen per a ponderar la força de les connexions entre neurones. Per aquest motiu es va pensar que la cognició i la memòria serien millors models per a l'aprenentatge que els models basats en la IA simbòlica. Tanmateix, els sistemes intel·ligents basats en el connexionisme tampoc necessiten formar part d'un cos ni estar situats en un entorn real i, des d'aquest punt de vista, tenen les mateixes limitacions que els sistemes simbòlics.

D'altra banda, el 90 % de les cèl·lules del cervell no són neurones sinó les anomenades cèl·lules glials, que no solament regulen el funcionament de les neurones sinó que també posseeixen potencials elèctrics, generen ones de calci i es comuniquen entre elles, la qual cosa semblaria indicar que també representen un paper molt important en els processos cognitius. No obstant això, no hi ha cap model connexionista que inclogui aquestes cèl·lules; per

tant, en el millor dels casos, aquests models són molt incomplets. Això fa pensar que l'anomenada singularitat, és a dir, futures superintel·ligències artificials que, basades en rèpliques del cervell, superessin per molt la intel·ligència humana en un termini d'uns vint anys, és una predicció amb poc fonament.

Una altra modelització bioinspirada, també compatible amb la hipòtesi del sistema de símbols físics, i no corpòria, és la computació evolutiva. L'èxit de la biologia evolucionant organismes complexos va fer que a primers dels anys seixanta alguns investigadors es plantegessin la possibilitat d'imitar l'evolució per tal que els programes d'ordinador, mitjançant un procés evolutiu, milloressin automàticament les solucions als problemes per als quals havien estat programats. La idea és que aquests programes, gràcies a operadors de mutació i encreuament dels «cromosomes» que modelen els programes, produeixin noves generacions de programes modificats de tal forma que les seves solucions siguin millors que les dels programes de les generacions anteriors. Atès que podem considerar que l'objectiu de la IA és la recerca de programes capaços de produir conductes intel·ligents, es va pensar que es podria fer servir la programació evolutiva per trobar aquests dins l'espai de programes possibles. La realitat és molt més complexa i aquesta aproximació té moltes limitacions, tot i que ha produït excel·lents resultats, en particular en la resolució de problemes d'optimització.

Una de les crítiques més fortes a aquests models no corporis es basa en el fet que un agent intel·ligent necessita un cos per poder tenir experiències directes amb el seu entorn en lloc que un programador proporcioni descripcions abstractes d'aquest entorn codificades mitjançant un llenguatge de representació de coneixements. Sense un cos, aquestes representacions abstractes no tenen contingut semàntic per a la màquina. No obstant això, mitjançant la interacció directa amb l'entorn, l'agent pot relacionar els senyals que percep mitjançant els seus sensors amb representacions simbòliques generades a partir del que ha percebut.

L'any 1965, el filòsof Hubert Dreyfus va publicar un article titulat «Alchemy and artificial intelligence» (Dreyfus, 1965) en el qual va afirmar que l'objectiu últim de la IA, és a dir la IA forta de tipus general, era tan inabastable com l'objectiu dels alquimistes del segle XVII que pretenien transformar el plom en or. Dreyfus argumentava que el cervell processa la informació de manera global i contínua mentre que un ordinador fa servir un conjunt finit i discret d'operacions deterministes, és a dir, aplicant regles a un conjunt finit de dades. En aquest aspecte podem veure un argument similar al de Searle, però Dreyfus, en articles i llibres posteriors, va usar també un altre argument basat en

Softbank Robotics

Una de les crítiques dels models no corporis d'IA es basa en el fet que un agent intel·ligent necessita un cos per poder tenir experiències directes amb el seu entorn. En la imatge, el robot humanoide Romeo desenvolupat por Softbank Robotics.

«PER MOLT SOFISTICADES QUE ARRIBIN A SER, LES INTEL·LIGÈNCIES DE LES MÀQUINES SERAN DIFERENTS DE LES NOSTRES»

IBM

La disponibilitat d'enormes quantitats de dades i l'accés a la computació d'altes prestacions per analitzar-les ha permès desenvolupar nous sistemes d'intel·ligència artificial com Watson, capaç de respondre a preguntes formulades en llenguatge natural. Segons la companyia IBM, que l'ha desenvolupat, Watson és capaç d'aprendre de cada experiència.

el paper crucial que el cos representa en la intel·ligència (Dreyfus, 1992). Va ser, doncs, un dels primers a advocar per la necessitat que la intel·ligència formi part d'un cos amb el qual poder interaccionar amb el món. La idea principal és que la intel·ligència dels éssers vius deriva del fet d'estar situats en un entorn amb el qual poden interactuar. De fet, aquesta necessitat de corporeïtat es basa en la fenomenologia de Heidegger, que emfatitza la importància del cos amb les seves necessitats, desitjos, plaers, penes, formes de moure's, d'actuar, etc. Segons Dreyfus, la IA hauria de modelar tots aquests aspectes per assolir l'objectiu últim de la IA forta. És a dir, que Dreyfus no nega completament la possibilitat de la IA forta, però afirma que no és possible amb els mètodes clàssics de la IA no corpòria.

ELS PROGRESSOS DE LA IA ESPECÍFICA ENS ACOSTEN A LA INTEL·LIGÈNCIA ARTIFICIAL GENERAL?

Pràcticament tots els projectes en IA s'han centrat a construir intel·ligències artificials especialitzades i els èxits assolits en només seixanta anys d'existència, i en particular durant l'últim decenni, són molt impressionants, principalment gràcies a la conjunció de dos elements: la disponibilitat d'enormes quantitats de dades i l'accés a la computació d'altes prestacions per a poder analitzar-les. Efectivament, l'èxit de sistemes com ara AlphaGo (Silver et al., 2016), Watson (Ferrucci, Levas, Bagchi, Gondek i Mueller, 2013) i els avenços en vehicles autònoms han estat possibles gràcies a aquesta capacitat per analitzar grans quantitats de dades. No obstant això, no hem avançat gens cap a la consecució d'IA general. De fet, possiblement la lliçó més important que hem après al llarg dels seixanta anys d'existència de la IA és que el que semblava més difícil (diagnosticar malalties, o jugar a escacs i a Go al més alt nivell) ha resultat factible i el que semblava més fàcil (comprendre el significat profund del llenguatge o interpretar una escena visual) encara no s'ha assolit.

L'explicació a aquesta aparent contradicció cal buscar-la en la dificultat de dotar les màquines de coneixements de sentit comú. El sentit comú és el requisit fonamental per aconseguir IA similar a la humana pel que fa a generalitat i profunditat. Els coneixements de sentit comú són fruit de vivències i experiències obtingudes interactuant amb el nostre entorn. Sense aquests coneixements no és possible una comprensió profunda del llenguatge ni una interpretació profunda del que capta un sistema de percepció visual, entre altres limitacions. 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 en 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 de les relacions entre ells.

> «EL SENTIT COMÚ ÉS EL REQUISIT FONAMENTAL PER ACONSEGUIR IA SIMILAR A LA HUMANA PEL QUE FA A GENERALITAT I PROFUNDITAT»

També necessitem nous algoritmes que, a partir d'aquestes representacions, puguin respondre, de forma robusta i eficient, a preguntes sobre pràcticament qualsevol tema. Finalment, com que necessitaran conèixer un nombre pràcticament il·limitat de coses, aquests sistemes han de ser capaços d'aprendre nous coneixements de manera contínua al llarg de tota la seva existència. En definitiva, és imprescindible dissenyar sistemes que integrin percepció, representació, raonament, acció i aprenentatge. Només combinant aquests elements dins de sistemes cognitius integrats podrem començar a construir IA general.

PASSAT RECENT I FUTUR A CURT TERMINI DE LA INTEL·LIGÈNCIA ARTIFICIAL

Entre les activitats futures, pensem que els temes de recerca més importants continuaran basant-se en allò que es coneix en anglès per *massive data-driven AI*,

és a dir, explotar la possibilitat d'accedir a quantitats massives de dades i poder-les processar amb maquinari cada vegada més ràpid per tal de descobrir relacions entre elles, detectar patrons i realitzar inferències i aprenentatge mitjançant models probabilístics com ara els sistemes d'aprenentatge profund (Bengio, 2009). No obstant això, aquests sistemes basats en l'anàlisi d'enormes quantitats de dades hauran d'incorporar en el futur mòduls que permetin explicar com s'ha arribat als resultats i conclusions que proposen, ja que la capacitat d'explicació és una característica irrenunciable en qualsevol sistema intel·ligent, ja que permet comprendre com funciona el sistema i avaluar-ne la confiabilitat. D'altra banda, també és necessari per corregir possibles errors de programació i detectar si les dades d'entrenament estan esbiaixades.

Cal saber si les respostes que ens donen són correctes per les raons correctes o a causa de coincidències que pot haver-hi en el conjunt de dades d'entrenament. Per això, un dels temes de recerca més importants en aprenentatge profund és dissenyar aproximacions interpretables d'aquests sistemes complexos d'aprenentatge. Una aproximació consisteix no només a entrenar el sistema d'aprenentatge profund sinó que, amb les mateixes dades, també s'entrena un altre sistema que el mimetitza usant una representació senzilla i transparent.

Un altre tema de recerca molt actual és la verificació i validació del programari que implementa l'algoritme d'aprenentatge. Això és especialment important en aplicacions d'alt risc com ara el pilotatge automàtic de vehicles autònoms. En aquests casos, necessitem una metodologia per a provar i validar que aquests sistemes d'aprenentatge automàtic assoleixen alts nivells de precisió. Una idea que s'està explorant actualment es coneix com a aprenentatge adversari (*adversarial learning* en anglès) i consisteix a entrenar un segon sistema d'IA que tracta de «trencar» el programari d'aprenentatge intentant trobar-ne els punts febles. Per exemple, en el cas del reconeixement visual, generant imatges que provoquin que el sistema prengui la decisió equivocada.

Waymo

El desenvolupament de tecnologies relacionades amb la intel·ligència artificial fa necessari analitzar els riscos que poden comportar. Per exemple, en el cas del pilotatge automàtic de vehicles autònoms necessitem una metodologia per a provar i validar que els sistemes d'aprenentatge automàtic assoleixen alts nivells de precisió.

ALTRES TEMES CLAU EN INTEL·LIGÈNCIA ARTIFICIAL

Altres àrees de la IA que continuaran sent objecte d'investigació extensiva són els sistemes multiagent, la planificació d'accions, el raonament basat en l'experiència, la visió artificial, la comunicació multimodal persona-màquina, la robòtica humanoide, la robòtica social, i les noves tendències en robòtica del desenvolupament que poden ser clau per a dotar les màquines de sentit comú. També veurem progressos significatius gràcies a les aproximacions biomimètiques per reproduir en màquines el comportament d'animals. Alguns biòlegs estan interessats en els intents de fabricar un cervell artificial el més complex possible perquè consideren que és una manera de comprendre millor l'òrgan i els enginyers, per la seva banda, busquen informació biològica per fer dissenys més eficaços.

«FINS I TOT SUPOSANT QUE FOS POSSIBLE DESENVOLUPAR PROGRAMARI COMPLETAMENT FIABLE, HI HA ASPECTES ÈTICS QUE ELS PROGRAMADORS HAN DE TENIR EN COMPTE»

Altres àrees importants d'interès per a la IA, i en particular per a la robòtica, són la ciència de materials i la nanotecnologia. Per exemple, per al desenvolupament de músculs artificials, cartílags artificials i sistemes sensorials com ara pells artificials.

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

Finalment, les aplicacions de la IA a l'art 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 música, arts plàstiques i narrativa, entre altres activitats creatives.

ELS RISCOS DE LA INTEL·LIGÈNCIA ARTIFICIAL, UNA REFLEXIÓ FINAL

Per molt intel·ligents que arribin a ser les futures intel·ligències artificials, en particular les de tipus general, mai seran iguals a les intel·ligències humanes, ja que, com hem argumentat, el desenvolupament mental que requereix tota intel·ligència complexa depèn de les interaccions amb l'entorn i aquestes interaccions depenen al seu torn del cos, en particular del sistema perceptiu i del sistema motor. Això, juntament amb el fet que les màquines no seguiran processos de socialització i culturització com els nostres, incideix encara més en el fet que, per molt sofisticades que arribin a ser, seran intel·ligències diferents de les nostres. El fet de ser intel·ligències alienes a la humana i per tant alienes als valors i necessitats humanes ens hauria de fer reflexionar sobre possibles limitacions ètiques al desenvolupament de la intel·ligència artificial. En particular, estem d'acord amb Weizenbaum (1976) quan diu que cap màquina hauria de prendre decisions de manera completament autònoma o donar consells que requereixin, entre altres coses, saviesa, producte d'experiències humanes, així com tenir en compte valors humans.

La IA es basa en programació complexa, i per tant necessàriament contindrà errors. Però fins i tot suposant que fos possible desenvolupar programari completament fiable, hi ha aspectes ètics que els programadors han de tenir en compte a l'hora de dissenyar-lo. Aquests aspectes ètics fan que molts experts en IA assenyalin la necessitat de regular el seu desenvolupament. Però, a més de regular, és imprescindible educar els ciutadans sobre els riscos de les tecnologies intel·ligents, dotant-los de les competències necessàries per a controlar-les en lloc de ser controlats per elles. Aquest procés de formació ha de començar a l'escola i tenir continuació a la universitat. En particular cal que els estudiants de ciència i enginyeria rebin una formació ètica que els permeti comprendre millor les implicacions socials de les tecnologies que molt probablement desenvoluparan. Només si invertim en educació aconseguirem una societat que pugui aprofitar els avantatges de les tecnologies intel·ligents minimitzant els riscos.

REFERÈNCIES

Bengio., Y. (2009). Learning deep architectures for AI. *Foundations and Trends in Machine Learning*, *2*(1), 1–127. doi: 10.1561/2200000006

Colton, S., Halskov, J., Ventura, D., Gouldstone, I., Cook, M., & Pérez-Ferrer, B. (2015). The Painting Fool sees! New projects with the automated painter. En *International Conference on Computational Creativity (ICCC 2015)* (pp. 189–196). Utah, UT: Brighma Young University.

Colton, S., López de Mántaras, R., & Stock, O. (2009). Computational creativity: Coming of age. *AI Magazine, 30*(3), 11–14. doi: 10.1609/aimag.v30i3.2257

Dreyfus, H. L. (1965). *Alchemy and artificial intelligence*. Santa Monica, CA: RAND Corporation.

Dreyfus, H. L. (1992). *What computers still can't do: A critique of artificial reason*. Cambridge, MA: MIT Press.

Ferrucci, D. A., Levas, A., Bagchi, S., Gondek, D., & Mueller, E. T. (2013). Watson: Beyond Jeopardy! *Artificial Intelligence*, *199*, 93–105. doi: 10.1016/j.artint.2012.06.009

López de Mántaras, R. (2016). Artificial intelligence and the arts: Toward computational creativity. En *The next step: Exponential life* (pp. 100–125). Madrid: BBVA.

McCulloch, W. S., & Pitts, W. (1943). A logical calculus of ideas immanent in nervous activity. *Bulletin of Mathematical Biophysics*, *5*, 115–133. doi: 10.1007/BF02478259

Newell, A., & Simon, H. (1976). Computer science as empirical inquiry: Symbols and search. *Communications of the ACM*, *19*(3), 113–126. doi: 10.1145/360018.360022

Searle, J. R. (1980). Minds, brains, and programs. *Behavioral and Brain Sciences*, *3*(3), 417–424. doi: 10.1017/S0140525X00005756

Silver, D., Huang, A., Maddison, C. J., Guez, A., Sifre, L., Van den Driessche, ... Hassabis, D. (2016). Mastering the game of Go with deep neural networks and tree search. *Nature*, *529*(7587), 484–489. doi: 10.1038/nature16961

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

Ramon López de Mántaras. Professor d'investigació i director de l'Institut d'Investigació en Intel·ligència Artificial del CSIC (Bellaterra, Espanya). Doctor en Física per la Universitat Paul Sabatier de Tolosa, *master of Science* en Informàtica per la Universitat de Califòrnia-Berkeley i doctor en Informàtica per la Universitat Politècnica de Catalunya. És membre numerari de l'Institut d'Estudis Catalans. Actualment investiga en raonament per analogia, en tècniques d'aprenentatge automàtic en robots humanoides i en intel·ligència artificial aplicada a la música, àrees en les quals ha publicat prop de 300 articles científics. Va publicar en 2017 el llibre de divulgació *Inteligencia artificial* dins de la col·lecció de llibres «Qué sabemos de» (Los Libros de la Catarata). ✉ mantaras@iiia.csic.es