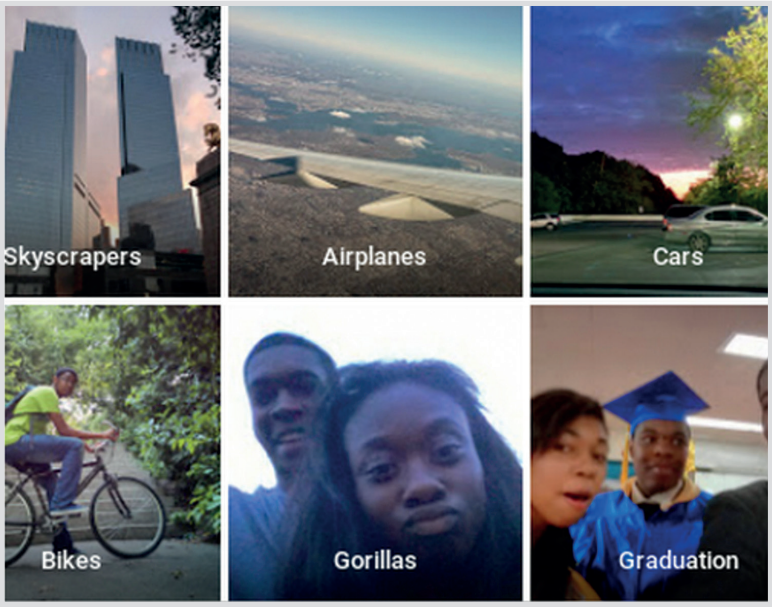
CAPCIT
Consell Assessor del Parlament sobre Ciència i Tecnologia

Butlletí del CAPCIT #18 · octubre 2020

Ètica i intel·ligència artificial



Els algoritmes de classificació d'imatges etiqueten el que «veuen» a partir de les dades d'entrenament que se'ls proporcionen. (Font: www.wsj.com/articles/BL-DGB-42522).

L'impacte de la intel·ligència artificial (IA) és cada vegada més gran i ha desencadenat arreu una mena de febre de l'or vinculada a aquestes tecnologies. Es creen nous laboratoris de recerca, apareixen centenars d'empreses emergents i creixen les inversions globals, particularment les de grans empreses de tecnologia digital. Es constata que la IA s'aplica cada vegada a més sectors, com ara el transport, les cadenes de producció i logística o el món financer. Les consultories expertes competeixen en les seves prediccions sobre l'impacte econòmic de la IA, i des de les institucions públiques i els governs es defineixen, aproven i executen plans estratègics per a adaptar les nostres societats al salt de la IA i evitar de quedar-se enrere. També a Catalunya, amb la recentment aprovada Estratègia d'intel·ligència artificial Catalonia.AI (http://smartcatalonia.gencat.cat/en/projectes/tecnologies/detalls/article/Catalonia-AI).

L'Estratègia d'intel·ligència artificial Catalonia.AI és un reconeixement a la ja llarga història de la IA a casa nostra. La recerca en IA a Catalunya es va iniciar a mitjan anys vuitanta amb la creació del Centre d'Estudis Avançats de Blanes, que va ser la llavor de l'actual Institut d'Investigació en Intel·ligència Artificial (IIIA) del Consell Superior d'Investigacions Científiques (CSIC), al campus de la Universitat Autònoma de Barcelona (UAB), i amb la creació del primer programa de doctorat en IA a la Facultat d'Informàtica (FIB) de la Universitat Politècnica de Catalunya (UPC). Força ràpidament la IA es va estendre per tot Catalunya amb grups de recerca a totes les universitats, fundats principalment per exalumnes de la FIB i de l'IIIA. Tot això va consolidar una comunitat científica que es va constituir formalment el 1994 com a Associació Catalana per a la Intel·ligència Artificial (ACIA) (https://www.acia.cat). Avui, l'ACIA aglutina la major part de la comunitat catalana de recerca en IA i organitza anualment el Congres Català d'Intel·ligència Artificial.

Podem afirmar que Catalunya ha esdevingut un nucli d'IA amb una projecció internacional de primer nivell. La comunitat catalana de recerca en IA ha aconseguit un ampli reconeixement internacional. Una mostra d'això és que Catalunya és actualment, amb 10 *fellows* (d'un total de 172 *fellows* a tot Europa), el quart territori europeu en nombre de *fellows* d'EurAI (European AI Association), només superat pel Regne Unit, Alemanya, França i Itàlia. El nivell de *fellow* (membre investigador) és el reconeixement més important, limitat a un màxim del 3% del total de membres.

Fins fa poc, tot i la qualitat de la recerca en IA a Catalunya, la seva transferència al teixit empresarial era més aviat escassa. Darrerament, però, diverses multinacionals han obert centres de recerca a Catalunya al voltant de la IA i, més important encara, també hi ha hagut un ràpid increment del teixit de pimes i empreses emergents interessades a introduir la IA en els seus models de negoci. De fet, segons Price Waterhouse Cooper, Barcelona es va identificar recentment com la quarta ciutat amb més futur tecnològic, just darrere de Singapur, Londres i Xangai. L'Estratègia d'intel·ligència artificial Catalonia.AI va identificar, el juny de 2019, 179 empreses en el sector de la IA a Catalunya. D'altra banda, *Digital Talent Overview 2019* (https://barcelonadigitaltalent.com/en/report/digital-talent-overview-2019/) situa la IA com el sector emergent amb més demanda a la ciutat de Barcelona. Es preveu un dèficit de 15.000 professionals de les tecnologies de la informació i la comunicació (TIC) a l'àrea de Barcelona. Aquesta dada està alentint l'adopció d'aquesta tecnologia per part de la majoria de les empreses. A mitjà termini part de la solució passaria per la creació de graus en IA a les universitats –de fet, ja hi ha diverses universitats que el curs 2021-2022 obriran graus en IA– i també per la formació professional. A més curt termini, caldria atraure talent que ara és fora i recuperar una part del talent format a les nostres universitats i centres de recerca que, al llarg dels darrers deu anys, ha

Autor: Ramon López de Mántaras. Fundador i ex-director de l'Institut d'Investigació en Intel·ligència Artificial (IIIA) del CSIC i Professor Adjunt de la Western Sydney University
https://www.iiia.csic.es/ i https://www.westernsydney.edu.au/research

hagut de marxar a altres països davant la manca d'oportunitats aquí.

Són bones notícies, però no podem negar els riscs derivats de l'aplicació de la IA. En aquest sentit, hi ha hagut iniciatives en els darrers anys per a entendre millor què comporta el desplegament de la IA i s'han elaborat marcs legals, codis de conducta i metodologies de disseny basades en valors ètics. Alguns exemples són els Principis d'Asilomar per a la IA (https://futureoflife.org/ai-principles/), la Iniciativa global IEEE sobre ètica de sistemes autònoms i intel·ligents (https://standards.ieee.org/industry-connections/ec/autonomous-systems.html), el consorci de la indústria tecnològica Partnership on AI per a beneficiar les persones i la societat (https://www.partnershiponai.org/) i el Reglament general de protecció de dades de la Unió Europea (https://gdpr-info.eu/). També creixen les aportacions sobre els riscs de la IA i com gestionar-la, com el debat que va tenir lloc al CosmoCaixa (Barcelona) el març de 2017, el resultat del qual va ser la Declaració de Barcelona sobre el correcte desenvolupament i ús de la IA a Europa.[1]

A mesura que la IA esdevé més sofisticada, els perills potencials derivats del seu ús també augmenten. Cal ser proactius en lloc de reactius respecte d'aquests nous perills i identificar les implicacions ètiques de les nostres tecnologies al més aviat possible, en comptes de desplegar-les i després anar apedaçant els problemes que apareixen. És imprescindible un debat públic i interdisciplinari sobre les qüestions ètiques en IA, ja que la ciència sola no les pot respondre. Ens cal l'ajuda de la filosofia, l'ètica i el dret.

La intel·ligència artificial planteja moltes qüestions ètiques. Tot seguit exposem algunes consideracions que creiem que cal tenir en compte i, per a cada una, formulem algunes preguntes concretes. No pretenem donar respostes, ja que creiem que el nostre deure com a societat és debatre col·lectivament possibles respostes i, si cal, regular per a evitar o, almenys, minimitzar els problemes derivats de l'ús de la IA.

Recomanem també veure el número monogràfic de la revista *Idees*,[5] en què diversos experts analitzen algunes de les polítiques científiques i industrials relacionades amb la IA més rellevants del món, així com quin és el seu impacte en l'economia i el mercat laboral, o en àmbits concrets com el de les ciutats «intel·ligents» o la indústria militar, i els seus efectes geopolítics i ètics. Així mateix, s'endinsen en el paper de la matèria primera de la IA, és a dir, les dades, particularment en la importància de les dades obertes, i reflexionen sobre el futur de la IA.

1. Biaixos dels algoritmes

Els algoritmes d'aprenentatge automàtic aprenen de les dades d'entrenament que se'ls proporcionen. Així, aquests algoritmes reflectiran –o fins i tot amplificaran– els biaixos que contenen les dades d'entrenament. Per tant, sempre ens hem d'esperar que estaran esbiaixats (https://www.technologyreview.com/2016/06/24/159118/why-we-should-expect-algorithms-to-be-biased/) i fer tot el que calgui per a evitar o, si més no, minimitzar els biaixos.

Per exemple, si s'entrena un algoritme amb dades racistes o masclistes, les prediccions resultants ho reflectiran (https://www.propublica.org/article/breaking-the-black-box-how-machines-learn-to-be-racist?word=Trump). És el que en anglès es coneix com a «*garbage in, garbage out*». Ja hem vist com alguns algoritmes de classificació d'imatges han etiquetat les cares de pell fosca com a «goril·les», o com han etiquetat com a assecador de cabells el que té una dona a la mà, quan de fet és un martell i, sense haver canviat res, llevat de la cara de la dona per la d'un home amb barba, l'algoritme l'etiqueta com a martell. I hem vist algoritmes en càmeres fotogràfiques que mostren el missatge «Ha parpellejat?» quan la cara del subjecte té trets orientals. En el cas dels goril·les es va eliminar la classe «goril·la» del conjunt de classes de l'algoritme, però, òbviament, aquesta no és una solució general. Altres algoritmes que s'usen per a determinar la solvència econòmica en demanar un crèdit bancari o per a filtrar els currículums per a un lloc de treball, també han estat criticats per pràctiques discriminatòries, ja que, tot i que en principi s'havien anonimitzat el nom, la raça i el gènere, els currículums contenien informació que permetia que l'algoritme deduís aquestes dades a partir d'altres dades no anonimitzades, com el codi postal o la universitat de la persona candidata. Per a evitar aquestes qüestions, caldria fer un escrutini detallat de l'algoritme abans d'usar-lo. Però, com es pot fer quan són propietat privada d'empreses i no són accessibles al control públic? Com podem equilibrar l'obertura del codi dels programes informàtics i la propietat intel·lectual?

2. Transparència dels algoritmes

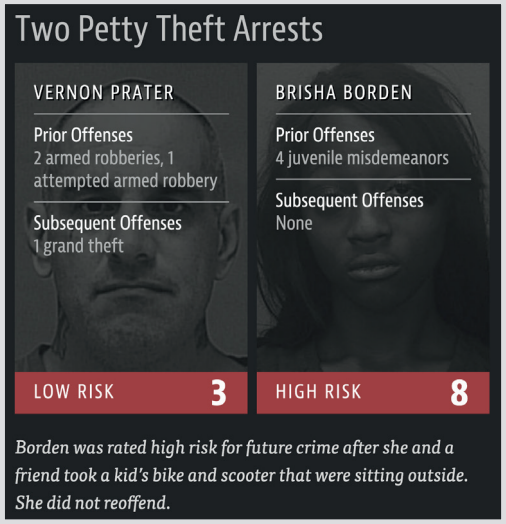
A més del problema que les empreses no permetin que els seus algoritmes siguin examinats públicament, hi ha el problema afegit que la majoria dels algoritmes actuals, basats en xarxes neuronals profundes, són com caixes negres fins i tot per als seus dissenyadors. És a dir que en molts casos no es pot fer una anàlisi de quines dades o característiques d'una descripció d'una tasca són les que han conduït a la solució proposada per l'algoritme. Per exemple, alguns algoritmes s'han utilitzat per a acomiadar treballadors, sense poder donar una explicació de per què el model predictiu indicava que havien de ser acomiadats (https://www.bloomberg.com/opinion/articles/2017-05-15/don-t-grade-teachers-with-a-bad-algorithm), i d'altres s'han fet servir per a negar la llibertat condicional a un condemnat, basant-se en la probabilitat de reincidència, també sense la capacitat d'explicar-ne els motius.

Com podem equilibrar la necessitat d'algoritmes més precisos amb la necessitat de transparència envers les persones afectades per aquests algoritmes? Si calgués, estaríem disposats a sacrificar precisió a canvi de transparència, tal com estableix el nou Reglament general de protecció de dades d'Europa? (https://arxiv.org/pdf/1606.08813v1.pdf). Tenint en compte que és probable que els éssers humans no siguem sempre conscients dels motius que ens han dut a prendre certes decisions o fer certes actuacions, hauríem d'exigir que les màquines donin explicacions que nosaltres no som capaços de donar?

3. Supremacia dels algoritmes

Una preocupació lleugerament diferent és la següent: si comencem a confiar en algoritmes per a prendre decisions, qui ha de tenir la darrera paraula en les decisions importants? Nosaltres o els algoritmes?

Com ja s'ha esmentat, alguns algoritmes s'usen per a avaluar la probabilitat de reincidència d'un con-



El perill de l'esbiax en els algoritmes d'aprenentatge automàtic es veu reflectit en l' estudi de ProPublica sobre l'ús d'aquests als EUA en les sentències contra criminals. (Font: https://www.propublica.org/)

demnat i decidir sobre la seva llibertat condicional, i fins i tot n'hi ha que determinen penes de presó (https://www.nytimes.com/2017/05/01/us/politics/sent-to-prison-by-a-software-programs-secret-algorithms.html). Tenint en compte que les decisions de qualsevol persona –i, per tant, també dels jutges– poden estar condicionades per nombroses circumstàncies, ja hi ha qui planteja que els jutges haurien de ser substituïts per algoritmes (https://www.technologyreview.com/2017/03/06/5361/how-to-upgrade-judges-with-machine-learning/). No obstant això, un estudi de ProPublica va trobar que, als Estats Units, un d'aquests algoritmes sobre sentències estava molt esbiaixat contra els afroamericans (https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing). En particular, després d'analitzar-lo es va poder detectar que l'algoritme usava dades sobre els coneguts d'un acusat que un jutge no hauria acceptat com a prova.

Hem de permetre l'ús de la IA en aquests casos? Els algoritmes IA haurien de tenir algun paper al món de la justícia o en altres àmbits que poden marcar profundament les nostres vides?

4. Notícies falses i vídeos falsos

Una altra preocupació ètica sorgeix al voltant del tema de la desinformació. L'aprenentatge automàtic s'usa per a determinar quin contingut s'ha de mostrar a diferents públics. Cal tenir en compte que la publicitat és la base del model de negoci de la majoria de plataformes de xarxes socials i que el «temps de pantalla» s'usa com a mesura típica de l'èxit d'una publicitat. D'altra banda, molta gent se sent atreta pels titulars sensacionalistes i per les històries esbiaixades i inflamatòries que es difonen de manera viral. Sabent això, és molt preocupant que la IA permeti crear vídeos falsos que cada cop costa més de distingir dels reals.

Per exemple, un estudi recent va mostrar que les notícies falses es propagaven més ràpidament que les notícies reals. Les notícies falses tenien un 70% més de probabilitats de ser repiulades que les reals (http://ide.mit.edu/sites/default/files/publications/2017%20IDE%20Research%20Brief%20False%20News.pdf). Per això, molts intenten influir en les eleccions i les opinions polítiques mitjançant notícies falses, com en el conegut cas de Cambridge Analytica.

Com podem frenar la difusió d'informació falsa i qui podrà decidir quines notícies es consideren certes? Si sabem que es poden falsificar vídeos i si en el futur pràcticament no els podrem distingir dels reals, haurien de ser acceptables com a proves en un judici? Com afectarà la IA a la geopolítica mundial? La IA reforçarà la democràcia o la posarà en perill?

5. Armes letals autònomes

La construcció o possessió d'armes letals autònomes aviat ja no estarà només a l'abast dels governs de les grans potències militars. Un vídeo de ficció relativament recent, creat per un grup d'investigadors en IA i divulgat pel Future of Life Institute (https://futureoflife.org/2017/11/14/ai-researchers-create-video-call-autonomous-weapons-ban-un/) ens alerta sobre la possibilitat que un eixam de drons autònoms es puguin usar per a atacar i eliminar persones amb característiques físiques concretes. Més de 4.500 investigadors en IA i robòtica hem signat una carta oberta en què demanem la prohibició de les armes autònomes (veg. https://futureoflife.org/open-letter-autonomous-weapons/?cn-reloaded=1). A la carta, entre altres aspectes, argumentem que, a diferència de les armes nuclears, les armes letals autònomes no requereixen matèries primeres costoses ni difícils d'obtenir, de manera que esdevindran omnipresents i barates per a la producció massiva a tots els països, en particular als que són potències militars. Només serà qüestió de temps que acabin al mercat negre i en mans de terroristes, dictadors que desitgin atemorir la població, senyors de la guerra que vulguin perpetrar una neteja ètnica, etc. Les armes autònomes són ideals per a finalitats com ara assassinar, desestabilitzar nacions, sotmetre poblacions i eliminar selectivament grups ètnics concrets.

Sobre quines bases hauríem de prohibir aquest tipus d'armes? Si es prohibeixen, com podem assegurar-nos que no es desenvoluparan igualment en clandestinitat?

6. Vehicles autònoms

Waymo, Uber, Tesla i gairebé tots els grans fabricants de vehicles motoritzats estan invertint molt per tenir un bon posicionament en l'àmbit dels vehicles amb un alt nivell d'autonomia. Els avenços són ràpids, encara que menys del que ens volen fer creure, però moltes qüestions ètiques i tècniques continuen sense resposta (https://www.vox.com/future-perfect/2020/2/14/21063487/self-driving-cars-autonomous-vehicles-waymo-cruise-uber). Per exemple, el sistema de visió artificial d'un Tesla no va detectar un camió, el Tesla el va envestir i l'ocupant del Tesla va morir (https://www.bloomberg.com/news/articles/2018-03-31/tesla-says-driver-s-hands-weren-t-on-wheel-at-time-of-accident). I un vehicle autònom d'Uber va atropellar mortalment un vianant que travessava el carrer i que el vehicle tampoc no havia detectat (https://www.nytimes.com/interactive/2018/03/20/us/self-driving-uber-pedestrian-killed.html). Tot i que en

En un article publicat en un número especial sobre IA a la revista *Idees*,[3] la professora Torras proposa crear cursos d'ètica per a tecnòlegs basats en la ciència-ficció i dona diversos exemples d'iniciatives existents al món. Aquest contingut és adequat per a un curs sobre ètica en robòtica social i intel·ligència artificial que tracti de com dissenyar l'assistent «perfecte», la importància de l'aparença de robots i la simulació d'emocions per a l'acceptació dels robots; del paper dels programes d'IA al lloc de treball i en ambients educatius; del dilema entre la presa de decisions automàtica i la llibertat i la dignitat humanes, i de la responsabilitat civil relacionada amb la programació d'una «moral» en els robots.[4]

ambdós casos hi havia conductors humans asseguts al volant, no van poder aturar el cotxe a temps. La tecnologia actual encara no és prou madura per a permetre que els sistemes de conducció automàtica tinguin el control absolut del vehicle en qualsevol lloc i qualsevol situació, és a dir, el nivell màxim d'autonomia (nivell 5). Recentment el conseller delegat (CEO) de Volkswagen va dir que el nivell 5 d'autonomia potser no s'assoliria mai (https://www.thedrive.com/tech/31816/key-volkswagen-exec-admits-level-5-autonomous-cars-may-never-happen). En qualsevol cas, a més de resoldre molts problemes tècnics, també caldrà respondre preguntes sobre qui hauria de ser responsable quan es produeixin accidents. El fabricant del vehicle?, els enginyers que van desenvolupar el programari, si és que va fallar? El fabricant dels sensors, si és que van fallar? O, en el cas del nivell 4, la persona asseguda al volant que no ha estat prou atenta? En el cas, poc probable, que un vehicle autònom no tingui cap altra opció que triar entre atropellar un vianant o xocar, què hauria de decidir l'algoritme que pilota el cotxe? Finalment, un cop els vehicles amb conducció autònoma siguin demostrablement més segurs que els conductors humans, hauríem de prohibir la conducció humana?

7. Privadesa *versus* vigilància

La presència generalitzada de càmeres de seguretat connectades amb algoritmes de reconeixement facial és una font de problemes ètics en relació amb la vigilància massiva de la ciutadania. Abans que es fes ús del reconeixement facial, les càmeres que fa anys que hi ha tant a llocs públics com privats no eren realment un problema greu per a la privadesa, ja que era impossible que els humans que estaven a càrrec de les càmeres observessin grans quantitats d'hores de filmació contínuament i només s'analitzaven les filmacions per motius ben definits relacionats amb qüestions de seguretat. En canvi, amb el reconeixement facial, els algoritmes poden analitzar molt ràpidament grans quantitats de metratge de forma contínua les vint-i-quatre hores del dia. Per exemple, és ben conegut que a la Xina s'usen càmeres per al control massiu dels ciutadans. A molts altres països, inclosa Espanya, hi ha la previsió que les forces de l'ordre usin el reconeixement facial. Alguns policies xinesos fins i tot han rebut ulleres de reconeixement facial que poden donar-los informació en temps real sobre els vianants que es troben pel carrer (https://www.theverge.com/2018/2/8/16990030/china-facial-recognition-sunglasses-surveillance). En canvi, en algunes ciutats dels Estats Units n'han prohibit l'ús en llocs públics i la Comissió Europea ha revelat que està considerant de prohibir-ne l'ús a les zones públiques durant cinc anys. Som a punt de fer realitat el panòptic de Jeremy Bentham a escala massiva? S'han superat les premonicions de George Orwell? Hauria d'haver-hi regulacions contra l'ús d'aquestes tecnologies?

Reflexions finals

No pretenem donar respostes, sinó més aviat contribuir a un debat necessari sobre les implicacions ètiques de la IA. Ara bé, el que sí que podem afirmar és que considerem imprescindible que es posin en marxa iniciatives educatives, tant en estudis secundaris com universitaris, en relació amb les qüestions ètiques al voltant de la IA. A la universitat ja s'ha començat a fer, però no és ni de bon tros una pràctica generalitzada. L'ensenyament de l'ètica en IA per a estudiants de carreres tècniques hauria de fer-se des d'un enfocament molt més pragmàtic que teòric. Els estudiants han de ser conscients de les implicacions ètiques i socials de la seva futura activitat professional i se'ls han de proporcionar eines per a analitzar i debatre aquestes qüestions. Les persones tenim conjunts de valors múltiples i sovint conflictius. L'objectiu no és unificar les opinions dels estudiants entorn d'un conjunt de normes, sinó augmentar la consciència i les habilitats per a pensar i discutir. Per exemple, la professora d'IA a la Universitat de Harvard Barbara J. Grosz defensa la integració de l'ètica en l'educació informàtica.[2]

La investigadora de l'Institut de Robòtica i Informàtica (UPC-CSIC) Carme Torras defensa que els graus universitaris tècnics s'haurien d'obrir a les humanitats perquè els estudiants siguin conscients de qüestions compromeses a què poden haver d'enfrontar-se i aprenguin a reflexionar i debatre sobre aquests temes. I es pregunta: s'hauria de fer complir la confiança pública en els robots i la IA? Si és així, com? És admissible que els dispositius estiguin dissenyats per a generar addicció? S'hauria d'excloure activament la possibilitat d'engany en el moment del disseny? S'hauria d'utilitzar la IA per a controlar les persones? La presa de decisions automàtica podria minar la llibertat i la dignitat de les persones? És acceptable que els robots es comportin com a substituts emocionals? Si és així, en quins casos? Es podrien usar robots com a terapeutes per a discapacitats mentals? En quina mesura haurien de ser adaptables/ajustables els programes i els robots? Hi ha límits a la millora humana per la tecnologia?

Referències

1. Luc Steels, Ramon López de Mántaras. «The Barcelona Declaration for the Proper Development and Usage of Artificial Intelligence in Europe». AI Communications, vol. 31(6): 485-494, 2018. DOI: 10.3233/AIC-180607
2. Barbara J. Grosz, David Gray Grant, Kate Vredenburgh, Jeff Behrends, Lily Hu, Alison Simmons, Jim Waldo. Communications of the ACM, agost 2019, vol. 62 (8): 54-61.
3. Carme Torras. «Ciència-ficció: un mirall per al futur de la humanitat». *Idees*, núm. 48, 2020. (https://revistaidees.cat/monografics/intelligencia-artificial/)
4. Carme Torras. La mutació sentimental. Pagès Editors, col·lecció «Ciència-ficció», núm. 22, 2008. ISBN: 978-84-9779-635-4.
5. Ramon López de Mántaras, Carles Sierra, Pere Almeda (editors). «IA: els interrogants que ens interpel·len». *Idees*, núm. 48, 2020.

PARLAMENT DE CATALUNYA

DL: B 26460-2013