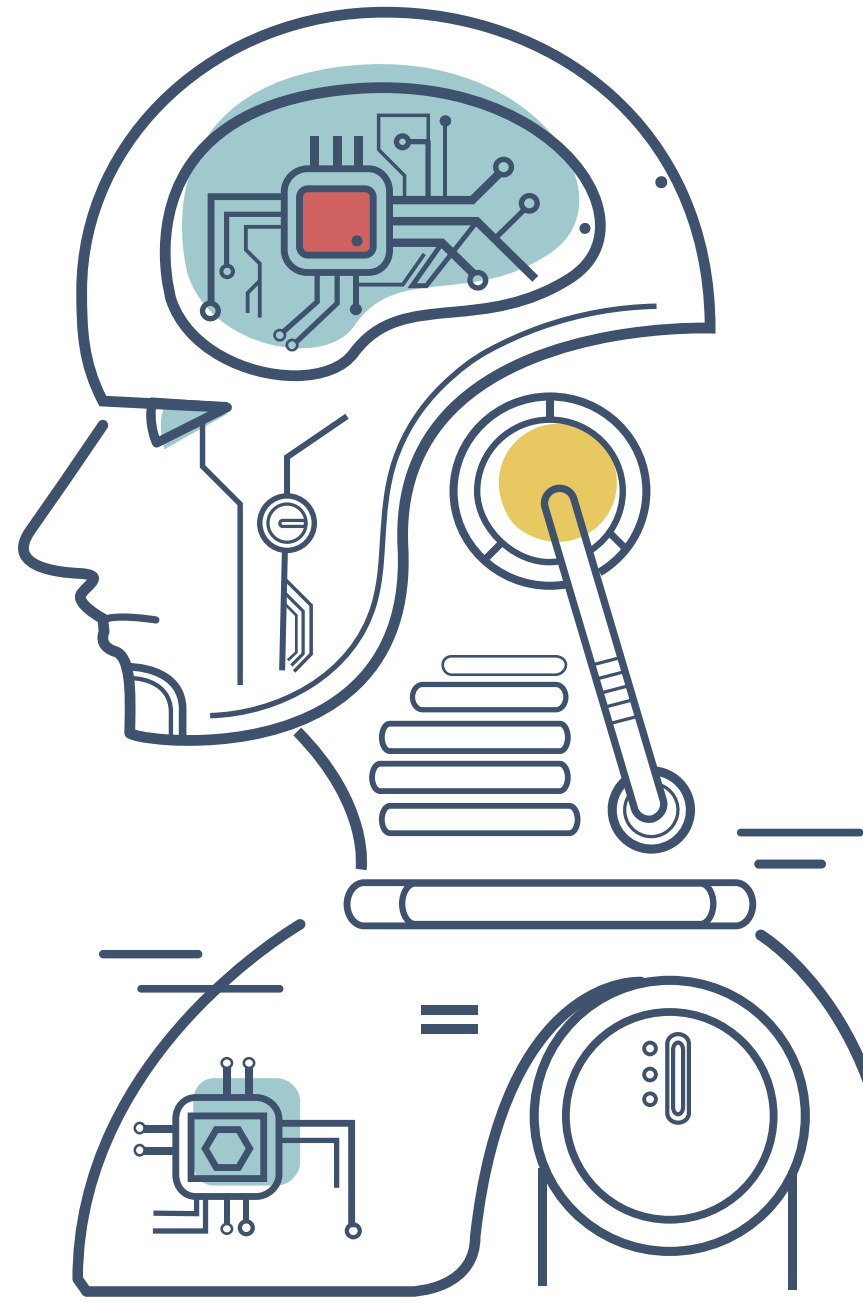
La Tribuna

La intel·ligència artificial ens deu una explicació



Les màquines intel·ligents no saben de lleis o costums: decideix a partir de les dades d'entrenament i pot ser que les seves decisions siguin políticament incorrectes

És incoherent demanar transparència a les administracions i tolerar que els ordinadors decideixin la nostra vida de manera opaca. Acceptar això darrer seria resignar-se a viure en una dictadura cibernètica

JOSEP DOMINGO FERRER
@diaridtarragona

Catedràtic distingit d'Enginyeria Informàtica URV

Josep Domingo és investigador ICREA-Acadèmia. Dirigeix la càtedra UNESCO de Privadesa de dades de la URV i és també el fundador i director del Centre de Recerca en Ciberseguretat de Catalunya (Cybercat). És membre de l'Institut d'Estudis Catalans i ha rebut la Medalla Narcís Monturiol al mèrit científic.

Aquests dies, arran del Congrés Mundial de Mòbils de Barcelona, sentim dir que els nous aparells incorporen cada cop més la intel·ligència artificial (IA). Les màquines intel·ligents poden ser un gran avantatge, però no estan exemptes de perills.

La IA va néixer els anys cinquanta del segle XX, quan els investigadors van començar a preguntar-se si les tasques intel·lectuals es podien automatitzar i si es podia aconseguir que les màquines pensessin. Els anys vuitanta va ser l'època dels sistemes experts: es formalitzava el coneixement d'experts humans per a problemes ben definits (diagnosi mèdica, diagnosi d'avaries, etc.) i s'introduïa aquest coneixement formalitzat en un ordinador. Per tant, les decisions que donava l'ordinador eren fàcilment explicables, atès que reproduïen el raonament de l'expert humà.

La intel·ligència artificial es comença a independitzar del raonament humà els anys 90, amb l'aparició de l'aprenentatge automàtic (*machine learning*). Si costava de formalitzar alguns problemes, es renunciava a la formalització i es feia que els ordinadors aprenguessin a partir d'exemples. Els algorismes de decisió dels ordinadors van esdevenir més difícils d'entendre i, per tant, les decisions, més difícils d'explicar a una persona. Val a dir que, en aquest estadi inicial, encara calia una intervenció humana important per preparar els exemples de manera que l'ordinador els pogués entendre.

Amb el segle XXI ha arribat l'aprenentatge profund (*deep learning*). Els algorismes d'aprenentatge profund, típicament xarxes de neurones artificials amb diversos nivells, poden aprendre de dades no estructurades com ara imatges que no necessiten ser preparades per humans. La xarxa neuronal pot aprendre d'extreure les característiques importants de les dades que hom li forneix. En contrapartida, el procés de decisió de l'aprenentatge profund encara és més difícil d'entendre i d'explicar a un humà que el de l'aprenentatge automàtic.

En els darrers cinquanta anys, els resultats de la intel·ligència artificial en els jocs han anat millorant de manera espectacular. Si els anys seixanta un ordinador podia guanyar un jugador humà ordinari a les dames i als escacs, a finals dels noranta va derrotar el campió mundial d'escacs. Amb el nou segle, l'aprenentatge profund ha derrotat el campió mundial del joc de go i ho ha fet sense que cap humà l'hagués d'entrenar. En l'aplicació de la IA als jocs, l'important és que l'ordinador guanyi i importa poc si les seves decisions són inesperades o inexplicables.

Però la intel·ligència artificial no s'aplica només a jocs, sinó a la concessió de préstecs, a la predicció financera, a les assegurances, als cotxes autònoms, a la diagnosi mèdica a partir d'imatges, etc. És en aquestes aplicacions que afecten la nostra vida que la societat vol saber per què l'ordinador ha decidit una cosa i no una altra: per exemple, per què ens ha denegat un crèdit. Aquesta exigència social d'explicabilitat ha estat recollida pel nou Reglament general de protecció de dades de la Unió Europea al seu article 22.

En efecte, si les persones no entenem les decisions automàtiques, és difícil que hi confiem. Malauradament, com acabem de descriure, com més intel·ligents són els algorismes de la IA, més difícil és explicarne el procés de decisió. Ara bé, l'explicabilitat és desitjable per a tothom, incloent-hi els enginyers informàtics que programen els algorismes i la societat en general.

Si els enginyers aconsegueixen entendre com i per què funciona l'aprenentatge profund, podran ajustar-ne millor els paràmetres, per tal que les decisions siguin més acurades i també més generalitzables, és a dir, que l'encertin encara que les dades que reben siguin una mica diferents de les dades d'entrenament.

Finalment, la societat vol que la intel·ligència artificial sigui explicable per poder descobrir decisions esbiaixades o discriminatòries. En efecte, la IA no sap de lleis, costums o acords polítics: decideix a partir de les dades d'entrenament i pot ser que les seves decisions siguin políticament incorrectes (per exemple, denegar més sovint els crèdits a un cert grup ètnic). És incoherent demanar transparència a les administracions i, en canvi, tolerar que els ordinadors decideixin la nostra vida de manera opaca. Acceptar això darrer seria resignar-se a viure en una dictadura cibernètica. Tenim dret a una explicació!