‘Machine learning’ i moralitat artificial

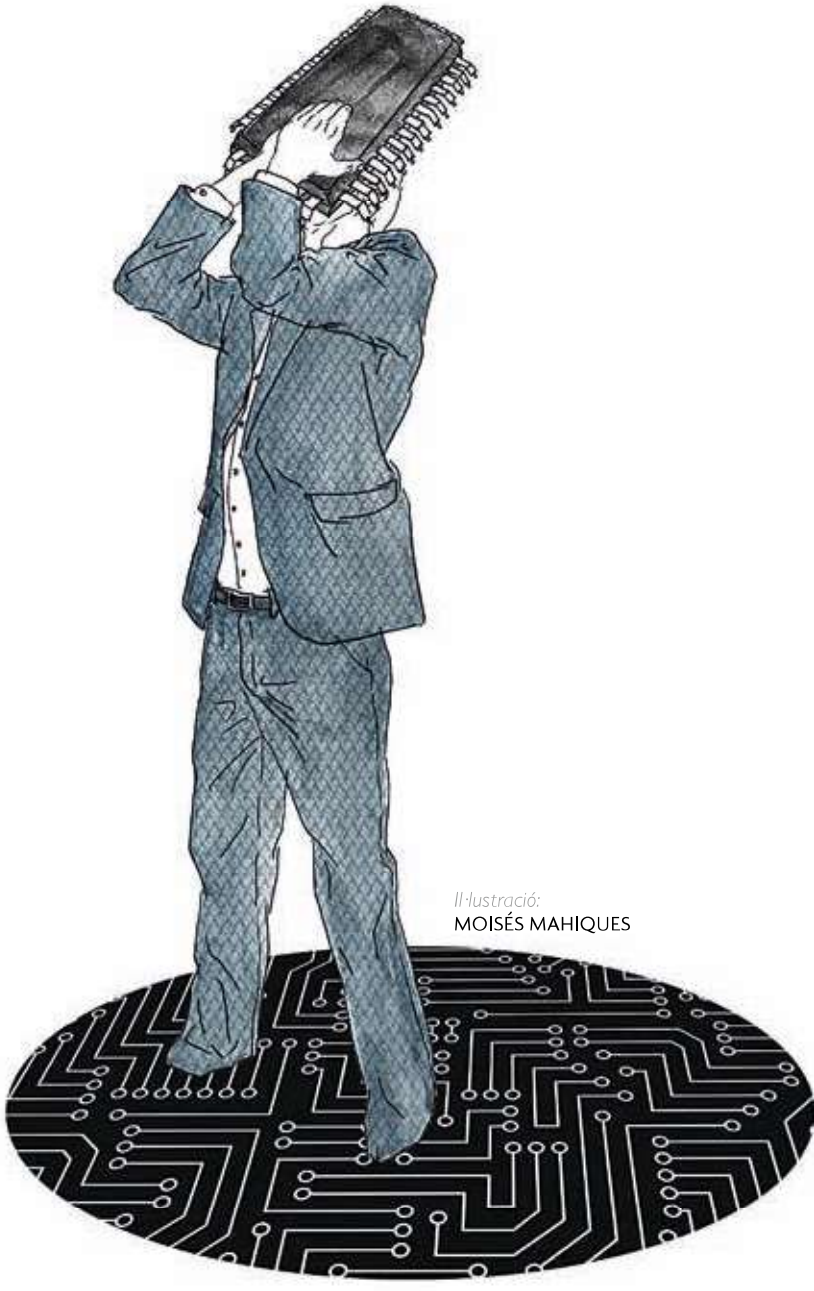
per **Pere Estupinyà**

Després de l’aparició en els anys quaranta dels primers ordinadors capaços de fer càlculs complexos, Alan Turing i altres científics de la computació es van preguntar si algun dia les màquines serien capaces de pensar de manera anàloga als humans. Va ser el naixement de la intel·ligència artificial i l’inici d’un vertiginós desenvolupament informàtic que generaria fites com la victòria en 1996 de l’ordinador Deep Blue sobre el campió mundial d’escacs Garri Kaspàrov.

Però la intel·ligència d’aquestes potentíssimes computadores no funcionava igual com la intel·ligència humana. Deep Blue basava el seu èxit en una programació molt precisa i un descomunal poder de càlcul que li permetia analitzar totes les situacions possibles davant de qualsevol moviment i esbrinar quina era la més reeixida probabilísticament. Una estratègia molt útil per a solucionar alguns tipus de problemes, però poc eficaç per a altres situacions en què les regles no estan tan definides com en els escacs. A aquella intel·ligència artificial li faltava una certa versatilitat, creativitat, intuïció…

La situació es va capgirar cap a 2012 quan van aparèixer els primers algorismes de computació que feien servir una estratègia diferent, el *machine learning* o aprenentatge automàtic, i que juntament amb el *big data* és el gran responsable de l’anomenada «nova onada de la intel·ligència artificial». El *machine learning* parteix d’un plantejament diferent: els programadors dissenyen uns algorismes per a, per exemple, reconèixer gats en fotografies, però després els comencen a donar milions de fotografies amb gats i sense perquè vagin comprovant si encerten i, quan cometen un error, modificar-se ells mateixos les seves línies de codi per a fer-se cada vegada més precisos.

El *machine learning* s’està incorporant amb força a qualsevol àrea on hi hagi presència massiva de dades que permeti aquest entrenament, com ara anàlisis genòmiques, econòmiques, gestió de transports o anàlisi del comportament humà a partir del nostre rastre digital. Dins del camp mèdic, un exemple paradigmàtic és l’anàlisi de radiografies, en què s’augura que en breu els algorismes d’intel·ligència artificial basats en *machine learning* cometran menys errors que els radiòlegs més experts. És inevitable preguntar-se quins impactes tindrà la intel·ligència artificial en els llocs de treball, fins a quin punt podran superar capacitats cognitives que consideràvem exclusives



Il·lustració: MOISÉS MAHIQUES

«És impossible plantejar tots els escenaris en què es podria trobar un cotxe autònom»

dels humans i quines decisions de la nostra vida quotidiana acabarem delegant en les màquines, per comoditat o perquè seran més llestes que nosaltres.

Els experts expliquen que els algorismes de *machine learning* seran excel·lents per a les funcions específiques que se'ls programi i que sens dubte ens superaran en tasques concretes, però que difícilment podran adquirir una «intel·ligència general humana» com la del nostre cervell programat per a infinitat de tasques al mateix temps. Així i tot, fa una mica d'angoixa, especialment amb l'afegit que, una vegada posats els algorismes a entrenar, en realitat perdem control sobre ells: van canviant i millorant-se per ells mateixos sense que nosaltres sapiguem què està ocorrent entre les seves línies de codi. És una caixa negra amb què, per exemple, un programa de pòquer ha après a fer catxa sense que ningú li hagi explicat com fer-ne, i del qual alguns temen que podrien aparèixer propietats, intel·ligències o comportaments emergents no previstos –ni desitjats.

Serà racista un programa que cerqui perfils en Linkedin per a un determinat lloc de treball? Et recomanarà una il·legalitat un algorisme que t'assessori en les teves finances amb l'objectiu de maximitzar els teus beneficis? Sona a ciència-ficció, però en realitat són escenaris molt plausibles, i per això diverses veus comencen a suggerir mesures de contenció a la intel·ligència artificial i a tenir molt en compte les consideracions ètiques en les decisions artificials.

Posem com a exemple el cotxe autònom que Iyad Rahwan, del Media Lab del MIT, utilitza per a preguntar-nos com a societat quines normes morals ha de seguir una màquina. El plantejament és el següent: d'aquí a uns anys hi haurà cotxes autònoms armats d'una visió perifèrica moltíssim millor que la nostra i amb capacitat d'anticipació més ràpida quan de sobte un nen travessi el carrer. Però si evitar atropellar aquell nen implica una volantada que provoqui aixafar un senyor gran que està esperant a la vorera, què ha de fer el cotxe? I si girar per a evitar el nen implica xocar contra una paret i posar en risc la integritat física del passatger? I si són quatre nens els qui creuen i hi ha un únic ocupant en el cotxe? Però més fonamental encara: qui ha de decidir?

En la nostra conducció actual prenem decisions instantànies sense temps a reflexionar, i en diem «accidents». Però en el futur aquestes decisions de mil·lisegons les prendrà el cotxe autònom basant-se en una sèrie d'instruccions. De nou, qui les estableix? Si ho fa el conductor clarament triarà protegir-se ell i si ho fan les companyies d'automòbils acabarà essent el mateix, perquè els compradors adquiriran el vehicle de la marca que més els protegeixi. El més lògic és pensar que els principis d'aquestes decisions morals s'acorden entre tota la societat, pensant que un dia pots ser vianant i un altre, passatger, i siguin comuns en tots els vehicles autònoms.

Us recomano anar al seu web i fer el test de la Moral Machine de Rahwan. L'escenari general sempre és el mateix: un cotxe autònom que porta passatgers pateix una fallada en els frens i ha de decidir de manera immediata entre dues situacions. Aquí alguns exemples: 1) atropellar, amb resultat de mort, dos xics i dues xiques atlètics que van pel seu carril travessant un pas de vianants o desviar-se un poc i atropellar dos homes i dues dones amb sobrepès que estan travessant el mateix pas de vianants davant seu; 2) atropellar dues xiques i dues àvies que estan creuant el pas de vianants en vermell o xocar contra una tanca i sacrificar dos homes i dos iaios que van dins del cotxe; 3) i si fossin quatre persones grans creuant el pas de vianants de manera completament legal i hi hagués quatre nens al cotxe autònom sense frens?; 4) i si pel pas de vianants creuen dos criminals i el cotxe porta quatre gats? I així fins a tretze situacions aleatòries

«El 'machine learning' s'està incorporant amb força a qualsevol àrea on hi hagi presència massiva de dades que permeti aquest entrenament»

destinades a comprovar quina importància relativa donem a salvar més o menys vides, a protegir els passatgers o els transeünts, les dones o els homes, joves o grans, persones amb un estatus social més o menys alt, o fins i tot humans o mascotes.

En acabar el test apareix un resum de les vostres preferències, que podreu contrastar amb la mitjana de totes les persones que han passat la prova. Interessant, de debò. També us preguntaran visions polítiques i religioses, ingressos o nivell educatiu, per als estudis que el grup de Rahwan està realitzant i ben aviat publicaran. Però de nou, l'aspecte més poderós conceptualment és que es fa impossible plantejar tots els escenaris en què es podria trobar un cotxe autònom. Nosaltres li podem donar una sèrie d'instruccions bàsiques, però la decisió final d'atropellar unes persones o altres la prendrà de manera autònoma una intel·ligència artificial a la qual demanarem que vagi aprenent i evolucionant pel seu compte, sense saber què està pensant ni com.

Pere Estupinyà. Escriptor i divulgador científic, Madrid. Presentador d'*El cazador de cerebros* (La 2).