# EL PERILL DE LA INTEL·LIGÈNCIA ARTIFICIAL PER A LA DEMOCRÀCIA

**JOSÉ MARICHAL**
Professor de Ciències Polítiques, California Lutheran University

L'any 2022 la revista *Nature* es feia ressò d'un estudi que presentava els resultats obtinguts per un grup d'investigadors que havien experimentat amb un *software* d'aprenentatge automàtic publicat per Collaborations Pharmaceuticals Inc., anomenat *MegaSyn*, que havia estat entrenat per identificar possibles fàrmacs. Havien sol·licitat a aquest programa que determinés possibles compostos tòxics que poguessin imitar la composició de l'agent nerviós VX. En menys de sis hores, l'algoritme va suggerir 40.000 armes biològiques potencials. Per descomptat, els investigadors van presentar aquest treball en un congrés sobre seguretat internacional amb l'objectiu de cridar l'atenció sobre els riscos potencials d'un mal ús de la Intel·ligència Artificial (IA). Tenint en compte aquest precedent, l'objectiu d'aquest article, des d'un punt de vista de la recerca en ciències socials, és extrapolar aquest cas a altres àmbits de la societat, i analitzar si podria passar una cosa semblant amb altres models lingüístics basats en IA, com ChatGPT, que avui està tant a l'ordre del dia.

La primera sorpresa per a molts dels que ens dediquem a analitzar els efectes sociopolítics de la tecnologia ha estat la velocitat inusitada amb què ChatGPT ha aconseguit que la IA aplicada al llenguatge passés de ser una eterna tecnologia de futur a ser un agent transformador del present. La seva irrupció ha deixat enrere la idea d'una Intel·ligència Artificial General (IAG), que és poc més que un maldestre assistent personal d'*smartphone* i que es limita a fer cerques quotidianes i de caràcter banal. ChatGPT3 és, clarament, una altra cosa; aspira a imitar amb una precisió sorprenent el pensament humà. Es tracta d'una maquinària de raonament que, per experiència, sabem que té les seves bondats, però que també genera els seus monstres.

La reeixida irrupció de ChatGPT ha empès altres empreses a llançar al mercat les seves pròpies versions d'aquestes eines, tot i desconèixer-ne les conseqüències potencials. No fer-ho implica quedar-se enrere i aquest és el pitjor pecat mercantil. Malauradament, tenim molts exemples dels problemes que això pot comportar. Per exemple, els investigadors Tristan Harris i Aza Raskin, del Center for Humane Technology, van mantenir una conversa amb el nou bot de la xarxa social Snapchat, My AI, en què es van fer passar per una nena de tretze anys a qui el seu xicot, de trenta-vuit anys, li havia proposat una escapada romàntica. La resposta de la IA va ser felicitar-la i suggerir-li maneres d'augmentar el romanticisme de la cita. I aquest és només un dels molts exemples que els usuaris han detectat i publicat a les xarxes.

Aquests fets no han passat desapercebuts. El març de 2023, un grup de personalitats rellevants del sector de la IA van signar una carta en què sol·licitaven la suspensió del desenvolupament de sistemes d'IA durant un període de sis mesos. Al cap de dos mesos, 27.535 persones s'havien adherit a la petició. En la mateixa línia, l'investigador informàtic Eliezer Yudkowsky va publicar un article d'opinió a la revista *Time* en què assenyalava que aquest termini era insuficient i sol·licitava una moratòria indefinida que tenia el suport d'acords internacionals; arribava a defensar fins i tot que els Estats Units tinguessin el beneplàcit internacional per «destruir militarment centres clandestins de processament de dades», en cas que una

nació o actor determinat vulnerés aquest acord. Amb tot, segurament, el senyal d'alarma més inquietant va ser la dimissió de Goeff Hinton –considerat el pare de la IA gràcies al seu treball capdavanter sobre xarxes neuronals a la dècada dels setanta– del seu càrrec a Google, per dedicar-se a explicar arreu del món el perill que la IAG caigui en mans equivocades i s'utilitzi per assolir altres objectius, com ara l'acumulació de poder.

El cert és que encara no es comprenen prou bé els perills que planteja realment la IA. Malgrat tota la sofisticació que s'oculta darrere del ChatGPT4, el model encara és lluny de sentir aquesta subjectivitat carnal que tan bé va retratar el poeta Walt Whitman en la seva invitació a «llegir aquestes fulles d'herba», en el poema homònim de 1860; la IAG no pot entendre l'amor, perseguir l'honor, sentir desesperació, soledat o autoenganyar-se. Tampoc no pot reflectir i avaluar el seu propi estat emocional. Quan demanem a la IA que actuï sobre el món, ho fa sense consciència ni reconeixement de si mateixa com a subjecte diferent d'un objecte. En termes heideggerians, la IA actua sobre el món, sense ser al món.

No obstant això, seria il·lusori pensar que la IA s'hagi d'assemblar a nosaltres per poder impactar-nos radicalment. Podem descartar, per exemple, que no serem nosaltres els que acabarem sent entrenats per la IA, i no a la inversa? I si, a causa dels seus efectes, acabem sent menys pensatius, menys reflexius i més dependents de la certesa en un món impulsat per l'optimització algorítmica? Segons el sociòleg Hartmut Rosa, a la nostra època impera una lògica d'acceleració que s'intensifica per la modernitat, si bé no en deriva exclusivament. Això ens provoca una voracitat per avançar, absorbir, produir i opinar, sense preguntar-nos quines són les raons més profundes que ens mouen a fer-ho.

Ateses aquestes circumstàncies, podem establir prioritats? És més preocupant una IA que sembra confusió i difon notícies falses que una altra que acumula poder? Segons la meva opinió, una IA que ens impedeix diferenciar la veritat de la ficció és ja, avui dia, un perill més gran per a la democràcia que una que estableix els seus propis objectius. Si el problema del segle XX va ser el relativisme de valors –la confrontació de la racionalitat il·lustrada per dues guerres mundials irracionals–, el del segle XXI és i serà el relativisme empíric: la incapacitat de saber amb certesa si allò que se sent, es llegeix o es veu és real. Si la crítica al relativisme moral sostenia que les posicions de valor eren reduïbles a gustos estètics, per a la IAG la realitat empírica seria només una opinió. Com decideixen els ciutadans democràtics entre les diferents realitats que se'ls presenten, quan aquestes realitats poden haver estat elaborades algorítmicament?

Ara com ara, una eina com OpenAI permet l'accés a l'API –Interfície de Programació d'Aplicacions– a qualsevol que vulgui entrenar el conjunt de dades d'aquesta «estúpida criatura» cap a les seves pròpies finalitats. Per tant, hem de concebre la necessitat de contractar un exèrcit de filòsofs perquè introdueixin en el model de la IA els fonaments rellevants de la filosofia política, o entrenar-la en els preceptes de l'utilitarisme i fer que produeixi resultats que permetin el major bé per al major nombre de persones. Segons l'enfocament aristotèlic de l'ètica de la virtut, desenvoluparíem la frònesi (saviesa cívica) a través de l'experiència. Així, doncs, podríem imaginar una IA que, amb la seva enorme capacitat de processament, pogués assolir la frònesi a velocitat de vertigen.

Però per més interessant que sigui un experiment de pensament d'aquest tipus, em resulta més interessant el procés de desaprenentatge de la IA. Presumiblement, si es pot entrenar una IA perquè millori contínuament a l'hora de prendre decisions, hi podria haver agents amb males intencions que manipulessin les dades d'entrenament per tal que una IA produís resultats incohe-

rents? Es pot doblegar la voluntat d'una IA? I, si arribés el cas, es podria tornar nihilista una IA i abandonar el projecte que se li hagi assignat?

Ja hem sentit a parlar d'exemples d'al·lucinació de ChatGPT3 quan les dades d'entrenament no li proporcionen prou informació sobre un tema. Això és per l'intent surrealista per part de GPT d'omplir les llacunes de coneixement amb les dades disponibles, però, i si l'al·lucinació en fos un tret característic i no pas un error? Es podria privar una IA de suficient informació o entrenar-la de manera que ignorés les seves dades d'entrenament originals i augmentessin els seus episodis al·lucinatoris? Seria una situació semblant a posar un pres en règim d'aïllament. Quin seria l'equivalent per a la IA el fet de tenir els llums encesos tot el dia i perdre la noció del temps i l'espai?

Per què algú hauria de voler entrenar així una IA? Semblaria contradictori, però n'hi ha prou de veure l'eficàcia amb què els governs –com els Estats Units– han utilitzat les xarxes socials –i els seus algoritmes de fidelització– per exacerbar la dissidència o una societat civil anèmica, incapaç d'oferir resistència a la voluntat de l'Estat, cosa que segons el politòleg James Scott és un dels atributs definitoris d'un Estat fallit.

Aquest darrer mecanisme de fracàs estatal és el que més em preocupa. En el context actual, coincideixen la necessitat creixent de tenir certeses i el debilitament de les societats civils. Al llibre *La condició humana* (Paidós, 1958), Hannah Arendt argumentava que la soledat era important per als ciutadans democràtics, ja que donava als individus la capacitat de la contemplació i que l'aïllament era, en canvi, un camí cap al totalitarisme. Quan estan aïllades, les persones se senten tan abandonades pels dels seus conciutadans que comencen a qüestionar-se a si mateixes i tot el que les envolta. Quan els ciutadans se senten aïllats són més vulnerables al totalitarisme. Arendt establia així la diferència amb la tirania, que és el govern de l'Estat motivat per la por. En un règim tirànic, la gent pot mantenir una vida privada que no sigui controlable per l'Estat, però en el totalitarisme, la ideologia impregna els ciutadans de manera que no hi ha distinció entre la vida pública i la privada. La IA es pot utilitzar per provocar més aïllament en els ciutadans augmentant la incertesa del món que els envolta?

En aquestes circumstàncies, ens podríem imaginar un poble inestable que va adoptant ideologies de la certesa i que veu l'Estat com un pare estricte disposat a imposar càstigs per preservar la llei. En aquest context, una IA entrenada per fer complir les normes adquireix més rellevància. Tindrem potser una IA dominadora i experta en les taules de la llei per produir resultats morals correctes? Per arribar a aquest punt ens hauríem d'aïllar prou els uns dels altres i dels nostres propis sistemes de significat. Així, doncs, hem de recuperar el sentit de nosaltres mateixos, tal com som, per evitar aquest destí i mantenir una democràcia sana i vivaç.