La intel·ligència artificial explicada als humans

La intel·ligència artificial explicada als humans

Jordi Torres

Pròleg de Mateo Valero

Plataforma Editorial

Primera edició en aquesta col·lecció: setembre de 2023

Plataforma Editorial
c/ Muntaner, 269, entl. 1a – 08021 Barcelona
Tel.: (+34) 93 494 79 99
www.plataformaeditorial.com
info@plataformaeditorial.com

Dipòsit legal: B 15481-2023
ISBN: 978-84-19655-58-5
IBIC: PDZ
Printed in Spain – Imprès a Espanya

Disseny i realització de coberta:
Sara Miguelena

Fotocomposició:
Grafime

El paper que s'ha fet servir per imprimir aquest llibre prové d'explotacions forestals controlades, on es respecten els valors ecològics, socials i de desenvolupament del bosc.

Impressió:
Sagrafic

A la meva família

Índex

Prefaci

A finals del 2022 hi va haver un punt d'inflexió en la nostra relació amb la intel·ligència artificial (IA) per causa, en gran manera, de l'aparició de diferents programes informàtics a l'abast de tots els usuaris. Aquestes IA permeten —a qualsevol persona amb accés a internet— generar textos i imatges que en molts casos és molt difícil saber si han estat creats per una IA o per un humà.

Això va animar un interessant debat públic: cap a on es dirigeix la IA i quines conseqüències pot acabar tenint per a la humanitat. Els mitjans de comunicació s'han fet ressò de com la IA forma part de pràcticament tots els aspectes de la nostra vida i del fet que canviarà el món de manera irreversible. Malgrat que com a societat encara ho estem assimilant, de moment no hi ha consens sobre on ens portarà la revolució de la IA en què ens trobem immersos.

Les opinions estan molt polaritzades —com succeeix darrerament en gairebé tot—. D'una banda, hi ha qui creu que la IA és una aliada que podrà aportar solucions als grans reptes que se li presenten a la nostra societat. D'altra banda, hi ha els que pensen que la IA és una enemiga de la humanitat,

potser per la influència que ha tingut la ciència-ficció i les distòpies amb les seves màquines amb superintel·ligència, generalment antropomòrfiques, capaces de superar l'ésser humà i rebel·lar-s'hi.

En tot cas, qualsevol eina poderosa pot ser beneficiosa o perjudicial, depenent de qui la utilitzi i amb quins fins. És a dir, tot i que la intel·ligència artificial té un gran potencial per millorar la nostra vida, el seu ús imprudent pot ser nociu i tenir un impacte negatiu en la humanitat.

En línies generals, regnen la inquietud i la confusió entre la població. Vegem-ho amb un exemple senzill: moltes persones, quan naveguen per internet o obren aplicacions al seu mòbil cada dia, no són conscients que estan fent servir una IA i que aquesta condiciona les seves accions. Aquest desconeixement els deixa totalment a la seva mercè.

Un altre exemple és l'aparició del *chatbot*, un servei d'IA gratuït amb una capacitat d'escriptura tan sofisticada que el text que produeix és inquietantment versemblant, tant que sembla escrit per un humà. L'arribada del ChatGPT, per esmentar el més conegut, ha posat en dubte la manera en què hem ensenyat i avaluat durant decennis els nostres estudiants.

El desconeixement genera confusió, temor, rebuig. Una de les causes principals d'aquest desconcert (entre el públic en general) és que es tendeix a utilitzar un llenguatge massa tècnic quan qui ho explica és un expert en la matèria. Tanmateix, quan qui ho explica és un divulgador, s'enfronta al desafiament de transmetre l'essència i la perspectiva del tema

amb un llenguatge clar, cosa que pot resultar difícil, si no impossible. I això, sens dubte, genera desassossec i, massa sovint, claudicació.

La IA ha revolucionat —i ho farà molt més— la forma en què interactuem amb el món, de manera que la seva comprensió resulta necessària per entendre com funcionen les relacions i la societat. Amb l'objectiu d'explicar de manera rigorosa, però accessible, el funcionament i l'impacte de la IA, he decidit escriure un llibre adreçat al públic no especialitzat.

Com veurem, tot i que el progrés de la IA és degut a tres vectors clau (els avenços algorítmics, la disponibilitat de grans quantitats de dades i la capacitat de computació), ha estat aquest darrer el que ha marcat el seu ritme de creixement. Per això, investigar en supercomputació resulta també de gran ajuda per poder explicar com ha evolucionat i com pot continuar evolucionant la intel·ligència artificial.

Aquesta proposta espera ser, precisament, una explicació rigorosa i accessible per a un lector sense coneixements tècnics i científics previs. Per a això, recorro a generalitzar els conceptes —sense faltar al rigor científic—, amb l'objectiu que el lector no expert en IA trobi la lectura amena i comprensible. S'aposta per un llenguatge planer per a la descripció dels conceptes fonamentals, alhora que es construeix un relat cronològic per facilitar-ne el seguiment —marcat per les efemèrides més útils per al relat—, i ajudant-se d'una selecció d'exemples tan familiars i propers al lector com sigui possible. El resultat d'aquest esforç de generalització ha

portat a definir quatre paradigmes en què es basa la intel·ligència artificial, que resultaran de gran ajuda per comprendre com hem arribat fins on som.

L'objectiu d'aquest llibre és fomentar la reflexió informada i conscient sobre l'evidència que ens trobem ja immersos, sense marxa enrere, en un nou paradigma coevolutiu en el qual la humanitat i la IA s'han embarcat conjuntament, tot gestant una interdependència i cohabitació que exigeixen respostes sense demora, perquè la IA no esperarà l'ésser humà.

Per això, aquest llibre intentarà respondre a vuit preguntes que cerquen sintetitzar les principals inquietuds al voltant d'aquest tema, ordenades de manera que componguin un relat de totes les etapes que ha viscut la IA, per, finalment, convidar el lector a reflexionar sobre quin tipus d'IA volem i on ens ha de portar:

1. Està la IA desplaçant l'ésser humà?
2. Com es va crear la primera IA?
3. Com una IA va començar a aprendre dels humans?
4. Com una IA va aconseguir aprendre per si mateixa?
5. Com una IA ha aconseguit ser creativa?
6. Podrà una IA arribar a pensar?
7. Ens ha de preocupar l'impacte de la IA en la seva forma actual?
8. Podem prescindir de la IA?

Crec sinceralment que tots tenim el dret d'entendre la revolució que suposa la IA, ja que és una de les més grans que

la humanitat ha experimentat, i a formar-nos-en la nostra pròpia opinió. D'aquesta manera, disposarem de les eines adequades a l'hora de prendre decisions sobre els potencials riscos i desafiaments associats a la IA, així com sobre les seves possibilitats i els seus avantatges. I podrem, a més, assegurar-nos que la IA evolucioni d'una manera responsable i sostenible, amb un impacte positiu en el futur de la humanitat.

5.
Com una IA ha aconseguit ser creativa?

El debat sobre la «creativitat» de la IA és damunt la taula des que es van popularitzar les primeres IA capaces de generar imatges sorprenents a partir d'un simple text. Per a molts, la creativitat és una qualitat intrínsecament humana que va més enllà de les arts i les ciències, com ho va demostrar Ferran Adrià a la gastronomia o Johan Cruyff al futbol. És a dir, entenem la «creativitat» com alguna cosa més que una creació, sinó també com una cosa innovadora que transforma i inspira altres a continuar sent creatius.

Podem avui dia considerar creativa una IA? És evident que la seva aplicació a l'art canviarà de manera contundent la naturalesa del procés creatiu. Encara més, deixarà de ser una eina d'ajuda a la creació per esdevenir una companya i, fins i tot, una agent creativa en si mateixa.

Tot i que hi ha qui no dubta a considerar creatives les IA com ChatGPT o DALL-E, altres veus, en canvi, opinen que no es poden tractar de «creatives», ja que tan sols «generen» respostes basades en l'aprenentatge propiciat amb milions de

dades creades, en darrera instància, per humans. És per això que només se'ls concedeix l'etiqueta de «IA generatives».

IA generativa

El 2021, OpenAI va llançar una nova IA per crear imatges a partir de text, batejada amb el nom de DALL-E, fent l'ullet al pintor Salvador Dalí. Aquest nou programari havia après a partir d'una gegantesca base de dades amb milions d'imatges descrites en text. Un any més tard, l'organització va llançar de manera gratuïta ChatGPT, l'arxiconegut bot conversacional que emula la lògica del pensament humà en la seva forma comunicativa.

Els avenços de les IA estan conquerint àmbits humans, com el llenguatge o la creació artística d'imatges, molt més ràpidament del que s'esperava. En els últims anys hem vist com les grans companyies tecnològiques han invertit gran quantitat de recursos en el desenvolupament de productes en aquesta direcció i estan reorientant la seva estratègia, llançant contínuament versions noves dels seus productes amb capacitats cada cop superiors, que ja no inclouen només textos de més qualitat, sinó que incorporen una IA multimodal, és a dir, capaços de treballar amb text, imatge i so alhora.

La fascinació que va despertar ChatGPT va ser tal que en només dos mesos des del seu llançament, a finals del 2022, sumava ja 100 milions d'usuaris únics. Per fer-se una idea de la magnitud del que representa, Instagram va trigar més de

dos anys a arribar a aquest mateix nombre d'usuaris. Aquestes xifres han convertit ChatGPT en l'aplicació de consum a internet de creixement més ràpid de la història fins ara. Potser la fascinació va ser deguda al fet que, per a nosaltres, els humans, el llenguatge és una finestra a la intel·ligència. De tota manera, sigui benvingut aquest interès sobtat per la IA de part de la societat.

Com s'arriba a un bot conversacional que sembla un humà

Tot va començar amb la fundació el 2015 de l'empresa Open AI per part d'un grup de líders tecnològics (Elon Musk entre ells, tot i que tres anys després va abandonar el projecte amb el pretext que l'empresa no investigava segons els objectius fundacionals, i el temps sembla que li ha donat la raó; ara és Microsoft la companyia que més hi està invertint). Des d'aleshores, OpenAI s'ha centrat a construir xarxes neuronals cada cop de mida més gran, entrenades al seu torn en supercomputadors cada cop més potents, els quals coneixem com a supercomputadors a gran escala.

La primera versió de xarxa neuronal que va presentar OpenAI va ser el model de llenguatge GPT, llançat el 2018, que utilitza 117 milions de paràmetres. Els models de llenguatge són un tipus de xarxa neuronal entrenats amb enormes quantitats de seqüències de lletres i paraules de diferents longituds, i fan servir un mecanisme diferent de la resta de

xarxes neuronals: els GPT estan dissenyats per prestar atenció a diferents parts d'una frase per tal de crear relacions entre elles. D'aquesta manera, rastreja on apareix cada paraula o frase dins d'una seqüència; gràcies a això pot «interpretar» el significat de les paraules segons el context. Aquestes probabilitats es calculen a partir d'una selecció de moltíssims textos en els quals el programa cerca amb quines altres paraules s'associa més sovint cada mot.

El 2019, OpenAI va presentar la següent versió d'aquest potent sistema de llenguatge natural, el GPT-2, compost ja per 1.500 milions de paràmetres. Aquesta versió nova ja estava configurada de tal manera que amb una breu indicació escrita (de només una o dues frases) fos capaç de generar una narració completa.

El maig del 2020 va arribar GPT-3, un sistema cent vegades més potent que l'anterior, amb més de 175.000 milions de paràmetres. El canvi va ser substancial. Els diferents GPT s'entrenen perquè siguin capaços d'endevinar quina ha de ser la paraula següent d'una frase. És a dir, el model genera un text paraula a paraula, tot executant-se iterativament l'algoritme de predicció una vegada i una altra per a cada paraula nova. La xarxa neuronal GPT-3 es va entrenar amb milers de milions de textos de diferents fonts d'internet, des de llibres i pàgines web fins a converses reals entre usuaris. Perquè el lector es faci una idea de la dimensió: la Viquipèdia sencera constitueix al voltant del 3 % del total d'informació amb què es va «alimentar» el nou programa.

L'entrenament és el següent: s'oculta una paraula del text

i s'executa la xarxa neuronal perquè la predigui. D'aquesta manera, l'esquema d'entrenament és equivalent al de l'exemple de la xarxa neuronal que classificava imatges segons si hi apareixia un gat o no. En aquest cas, sabem quina és la solució que busquem perquè és precisament la paraula que se li ha ocultat al programa. Finalment, es compara el valor que ha calculat la xarxa neuronal amb l'esperat per ajustar els valors dels paràmetres de la xarxa neuronal (vegeu el capítol 3).

Recordem que una interpretació simple del que contenen els milers de milions de paràmetres de la xarxa neuronal és la seva versió comprimida de tot el coneixement que se li ha mostrat per aprendre. És un procés similar al de comprimir un arxiu. Requereix dos passos: primer, la codificació per comprimir, durant la qual l'arxiu es converteix a un format més compacte, i després la descodificació a partir de la informació comprimida, en la qual s'inverteix el procés. És a dir, quan fem servir aquestes IA per generar text, en realitat estem descodificant i, per tant, la seqüència exacta de paraules que eren a les dades originals no es troba emmagatzemada tal qual, ja que la còpia comprimida només és una representació de la informació real. Tot i això, en descodificar, és possible obtenir una aproximació en forma de text gramatical equivalent. Això explica alguns casos en què les respostes de les IA són poc encertades, perquè en certa manera és inevitable que s'hagi perdut informació en el procés de compressió.

És important remarcar que aquestes IA generatives de l'estil de GPT tendeixen a ser adaptables, i això vol dir que

poden adquirir altres habilitats a més d'aquelles per a les quals van ser explícitament capacitades. Això és possible gràcies al seu entrenament generalista. GPT-3, per exemple, no només va aprendre a escriure un text d'aspecte realista, sinó que també va aprendre a generar un codi de programació acceptable, malgrat que no tenia la intenció explícita de ferho al principi.

A finals del 2022 es va llançar la versió oberta de ChatGPT (una versió millorada de GPT-3). Aquesta vegada, el programa es va centrar a utilitzar el contingut de les converses interactives entre persones. Recordem que el model del llenguatge GPT-3 només estava entrenat per predir la paraula següent en una seqüència de text, però era incapaç de comprendre'n el significat. A la versió nova es va millorar el procés d'aprenentatge amb la inclusió de comentaris humans i amb tècniques d'aprenentatge per reforç, però fent servir una retroalimentació amb intervenció humana al cicle d'entrenament. Per què? Perquè basar l'entrenament en textos extrets d'internet hauria tingut un efecte col·lateral no desitjat: junt amb la informació vàlida, GPT-3 hauria absorbit gran part de la desinformació i els biaixos que es troben a la xarxa. Per això, per reduir la quantitat d'informació errònia i de textos ofensius que produïa GPT-3, va caldre ajustar-la de forma «manual». Al final, el procés d'entrenament de les IA generatives requereix que la mà humana hi sigui molt present.

Aquests models segueixen evolucionant a mesura que n'escrivim la història. Poc després de llançar el primer ChatGPT, OpenAI anunciava una versió nova, GPT-4, que ad-

met com a entrada no només text sinó també imatges. Aquesta vegada, la companyia no ha fet públic ni el seu entrenament ni detalls sobre els paràmetres o requeriments computacionals. En tot cas, representa una millora de la versió anterior, tot i que encara és propens als mateixos tipus de problemes de veracitat.

Poder transformacional de les IA generatives

ChatGPT encara no és capaç de llegir un llibre (entenent per llegir la facultat de comprendre'n el contingut). L'aproximació actual de qualsevol de les IA disponibles es basa en la representació de probabilitats intentant endevinar quines paraules tendeixen a concórrer en una frase o context. És a dir, es generen textos que semblen escrits per humans, però no vol dir que la IA tingui coneixements del tema ni que hagi entès el text. És una cosa similar al predictor de text de WhatsApp, que ens suggereix paraules per completar el missatge que estem escrivint.

Un dels aspectes més preocupants és la manca de veracitat i els biaixos que esmentàvem a l'apartat anterior. Moltes vegades la IA és capaç de respondre amb informació falsa com si fos certa, sigui perquè les dades d'entrenament no estan actualitzades o perquè en el procés de codificació i descodificació s'ha perdut informació.

No s'ha de perdre de vista que la xarxa neuronal no és més que un model limitat del món format amb els valors

dels seus paràmetres, i no pas un model del món real. De moment, no s'ha trobat la manera d'entrenar models amb dades extretes d'internet sense absorbir el que es coneix com la brutícia de les dades, és a dir, les boles i continguts ofensius, entre altres. Per això, igual que hem vist en el cas de GPT-3, l'única solució que hi ha fins ara és que operadors humans filtrin la informació a mà.

Malgrat totes aquestes limitacions, l'onada de les IA generatives no ha fet més que començar, i ChatGPT representa només el primer exponent dels models nous que aviat seran presents en tots els aspectes de les nostres vides. Les seves possibilitats són gairebé infinites: des de xatejar fins a generar documents sofisticats o senzillament servir d'inspiració. Com ja ha succeït amb els traductors automàtics, les aplicacions de creació de text seran habituals als nostres dispositius i passaran a integrar-se al programari que fem servir quotidianament en les nostres rutines productives, tant en l'àmbit domèstic com en l'empresarial i educatiu.

Els gegants tecnològics habituals s'han llançat a una cursa per desenvolupar IA generatives no només molt més potents i eficients, sinó també específiques per a àmbits determinats, és a dir, entrenades amb dades personalitzades (per exemple, medicinal o empresarial). Aviat es convertiran en les primeres expertes en l'àrea en què hagin estat entrenades.

Un darrer apunt sobre aquesta cursa per la IA: la proliferació de desenvolupadors ha generat també tensions entre la comunitat de codi obert IA i les empreses privades respec-

te de l'exclusivitat dels codis d'IA. Els primers advoquen a favor d'unes IA generatives com a instrument de creativitat i innovació obertes, mentre que els segons en defensen la privatització.

La IA basada en força bruta

Però com hem arribat fins aquí? Ja hem dit que l'any 2012 va ser un punt d'inflexió per a l'adopció de les xarxes neuronals, en especial gràcies a l'equip de la Universitat de Toronto i la seva revolucionària participació en la competició d'ImageNet.

Aviat es va veure que els mètodes d'aprenentatge de les xarxes neuronals podien aprofitar magníficament tècniques de paral·lelisme, és a dir, utilitzar diversos xips acceleradors de forma simultània per reduir el temps d'entrenament de les xarxes neuronals. I, finalment, els supercomputadors a gran escala van permetre accelerar això encara més, tot interconnectant una gran quantitat de màquines amb diversos xips acceleradors cadascuna.

Sens dubte, el paral·lelisme és una tècnica cabdal en la supercomputació a gran escala. Posem com a exemple la xarxa neuronal de Google per a traducció multilingüe. Consisteix en una xarxa amb sis-cents mil milions de paràmetres, és a dir, una capacitat de computació equivalent a vint-i-dos anys si només es disposés d'un xip accelerador tipus TPU (com el que es va fer servir a AlphaZero). Però com que el sistema

de Google utilitza 2.048 xips d'aquest tipus en simultani, aconsegueix fer l'entrenament en només quatre dies.

L'últim any, les necessitats de computació per entrenar les IA generadores de text s'han multiplicat per dos cada tres o quatre mesos, de manera que les infraestructures amb gran capacitat de computació s'han revelat com a fonamentals. Avui dia és inconcebible pensar en un supercomputador a gran escala que no disposi d'un maquinari pensat per entrenar una IA. Un dels més recents és el MareNostrum 5 (ha entrat en funcionament durant l'any 2023), que inclou 4.480 xips acceleradors GPU d'última generació fabricats per Nvidia. És un dels nodes principals de la xarxa europea de supercomputació EuroHPC.

La gran capacitat computacional disponible actualment ha permès a la comunitat d'IA avançar els últims anys amb molta rapidesa i dissenyar xarxes neuronals cada cop més complexes, encara que això ha exigit augmentar la infraestructura de computació a nivells mai vistos. Estem immersos en el que podríem anomenar el paradigma de la IA basada en «força bruta». És a dir, algoritmes de milers de milions de paràmetres que necessiten supercomputadors a gran escala per ser entrenats amb quantitats ingents de dades. Uns recursos, per descomptat, a l'abast de ben pocs.

En resum,
com una IA ha aconseguit ser creativa?

- ChatGPT va ser una IA generativa orientada a la comprensió i generació de diàlegs per conversar amb els humans basada en un tipus de xarxa neuronal de milers de milions de paràmetres.
- Aquest tipus de xarxes pot adquirir habilitats molt més complexes d'aquelles per a les quals van ser capacitades a causa del seu entrenament generalista amb grans conjunts de dades.
- Per evitar absorbir tant com es pugui la desinformació i els biaixos que conté internet, s'ha millorat l'entrenament de les xarxes neuronals amb tècniques d'aprenentatge per reforç en el qual intervenen equips humans.
- L'onada d'IA generatives no ha fet més que començar. ChatGPT només representa el primer exponent d'unes IA generatives que aviat seran presents en tots els aspectes de les nostres vides i comportarà un impacte social sense precedents.
- La IA generativa està basada en el paradigma de força bruta, un escenari en què xarxes neuronals de milers de milions de paràmetres han de ser entrenades amb grans quantitats de dades, i per fer-ho es requereixen supercomputadors a gran escala.

La seva opinió és important.
En futures edicions, estarem encantats de recollir els seus comentaris sobre aquest llibre.

Si us plau, enviïn-los a través del nostre web:

www.plataformaeditorial.com

Per adquirir els nostres títols,
consulteu el vostre llibreter habitual.

«*I cannot live without books*».
«No puc viure sense llibres».
THOMAS JEFFERSON

Plataforma Editorial planta un arbre per cada títol publicat.