CIVICAi

PROTOCOL PER A UNA INTEL·LIGÈNCIA ARTIFICIAL CÍVICA*

Capacitació ciutadana per a una governança cívica de la IA

* El contingut i el text d'aquest protocol han estat concebuts i redactats per humans (nivell 1 AIAS). La seva versió final ha estat revisada per la IAg amb l'objectiu de corregir errades tipogràfiques, identificar possibles mancances i suggerir millores en el contingut i la claredat expositiva, principalment pel que fa al glossari de l'annex B (nivell 3 AIAS).

CIVICAi

RESUM EXECUTIU

El Protocol d'Intel·ligència Cívica examina la Intel·ligència Artificial Generativa (IAg) com una tecnologia transformadora que revoluciona la nostra comprensió de la intel·ligència, el llenguatge i la cognició[1]. El document explora quatre àmbits clau: el llenguatge, la intel·ligència, els riscos i reptes, i la governança de la IAg.

En l'àmbit del llenguatge, el protocol compara l'enfocament tradicional basat en regles gramaticals innates amb els models de llenguatge actuals que aprenen de grans volums de dades i estableixen relacions vectorials contextuals, una manera d'aprendre similar al processament cerebral humà.

Respecte a la intel·ligència, el protocol analitza profundament les capacitats emergents d'aquests sistemes quan assoleixen determinada complexitat, destaca la seva habilitat per adaptar recursos computacionals a diferents tasques (*test-time compute*), i proposa arquitectures híbrides que combinen enfocaments connexionistes i simbòlics per superar limitacions actuals.

El document identifica riscos a curt termini (biaixos, vulneracions de privadesa, desplaçament laboral, desinformació) i a llarg termini (singularitat tecnològica, desalineament d'objectius, pèrdua d'autonomia humana). També aborda preocupacions sobre sostenibilitat energètica i la dificultat creixent d'accedir a dades de qualitat per a l'entrenament dels sistemes d'IA avançats.

Finalment, proposa un marc de governança global amb un organisme internacional que integri governs, experts, societat civil i empreses, amb capacitat per registrar, monitorar i regular sistemes d'IA avançats, prevenir monopolis i implementar mecanismes de participació ciutadana directa. L'objectiu és democratitzar no només el coneixement sobre la IA sinó també els processos de decisió sobre el seu desenvolupament, per garantir que reflecteixi la diversitat de valors socials.

El protocol inclou dos annexos complementaris: un recull de preguntes i respostes sobre els temes tractats i un glossari de terminologia específica relacionada amb la IA.

[1] Christopher Summerfield (2025). These Strange New Minds: How AI Learned to Talk and What It Means. Nova York: Viking. 978-0-593-83171-7

CIVICAi

ÍNDEX

CIVICAi

INTEL·LIGÈNCIA ARTIFICIAL CÍVICA

Una de les tecnologies més representatives d'aquesta profunda transformació tecno-social[2], i que la converteix en veritablement revolucionària, és la Intel·ligència Artificial (IA), caracteritzada pel seu desenvolupament accelerat i el seu impacte transversal en tots els àmbits de la societat. L'adveniment dels sistemes generatius d'IA (IAg), entrenats a partir de grans volums de dades de text — grans models avançats de llenguatge (LLMs) — o de dades d'imatges, d'àudio i de vídeo, així com la propera arribada de la Intel·ligència General Artificial (AGI), representa un canvi revolucionari en l'evolució humana i en la nostra comprensió de la intel·ligència, el llenguatge i la cognició.

La revolució tecnològica de la IA no només transformarà les nostres eines i mètodes de treball, sinó que remodelarà la nostra comprensió de conceptes fonamentals com la intel·ligència i el coneixement. Com a membres de CIVICAi, assumim la responsabilitat d'enriquir el discurs públic i potenciar la comprensió social d'aquests canvis profunds i accelerats. El nostre objectiu és que, quan aquests canvis s'hagin consolidat, l'evolució subsegüent pugui integrar-se harmònicament en la societat i servir al bé comú.

L'estructura del protocol (Figura 1), amb els quatre apartats següents — centrats en el llenguatge, la intel·ligència, els riscs i els reptes, i la governança de la IAg — i el recull de preguntes i respostes proposades a l'annex A sobre aquests mateixos temes, han estat concebuts per proporcionar un mapa amb informació suficient perquè el lector navegui per les complexitats de la IAg. L'objectiu és examinar tant els matisos tècnics com les implicacions filosòfiques més àmplies relacionades amb la IAg, amb una exposició entenedora i el rigor conceptual indispensable. Aquest recorregut pels fonaments conceptuals, capacitats i reptes de la IAg ens permetrà establir les bases per a una participació ciutadana informada en el desenvolupament i regulació d'aquestes tecnologies transformadores.

En un moment en què la IA avança a un ritme sense precedents, aquest protocol vol servir com a eina d'empoderament per a una ciutadania que haurà de conviure i coevolucionar amb sistemes d'intel·ligència artificial cada vegada més sofisticats.

[2] En aquest context és important reconèixer que el fenomen de la IA, especialment la IAg i la potencial Intel·ligència General Artificial, constitueix una transformació que transcendeix el marc conceptual de les "revolucions industrials", històricament seqüencials. La IA no representa simplement una fase evolutiva de la industrialització, sinó possiblement l'inici d'una nova era transformadora comparable en magnitud i profunditat a les grans transicions històriques de la humanitat, com el pas a l'agricultura, la industrialització o la digitalització.

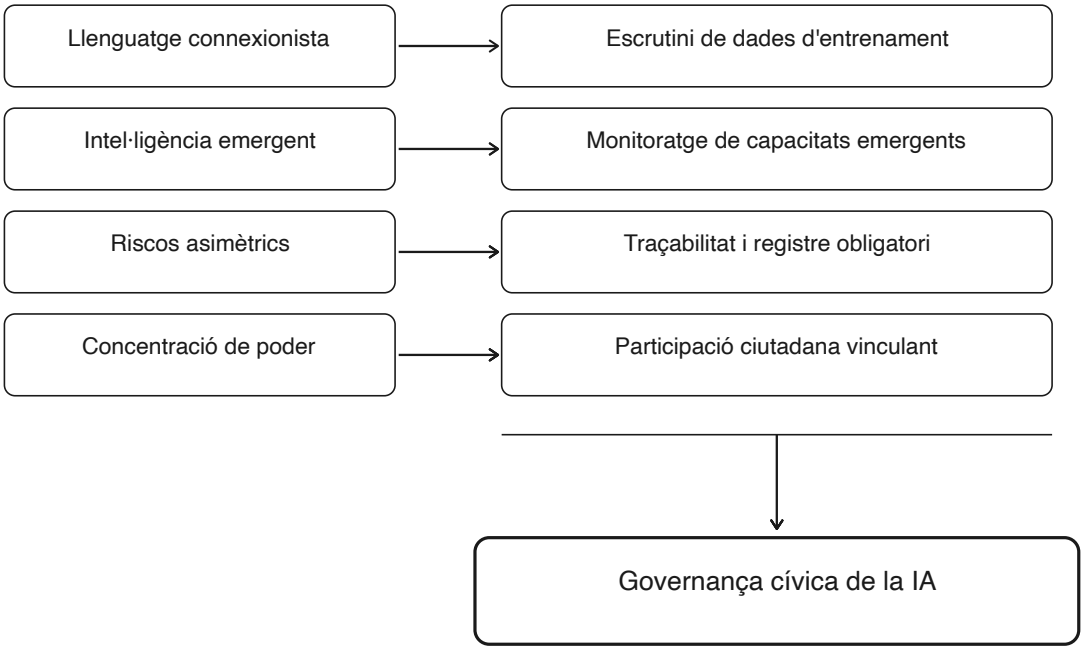

Figura 1. Estructura del Protocol: cada àmbit conceptual sosté una implicació de governança concreta; totes quatre conflueixen en la governança cívica de la IA com a punt d'arribada.

1. El llenguatge

Històricament, la nostra concepció de la intel·ligència ha estat profundament influenciada pel dualisme cartesià, on René Descartes postulava una separació estricta entre la ment i el cos[3]. Els inicis de la IA també varen ser influïts per la teoria de la gramàtica universal de Noam Chomsky, introduïda en la dècada dels 60, la qual proposava que la capacitat lingüística és innata (inherent) en el cervell humà — que hi està programada de forma natural. Segons aquesta teoria, l'aprenentatge lingüístic es fonamenta en estructures i regles innates existents en el cervell humà, fet que resulta contrari a l'aprenentatge basat en la identificació de patrons a partir de grans quantitats de dades[4]. Aquest context històric va influir en els primers passos de la IA, impulsant la denominada IA simbòlica, enfocada a simular la cognició humana a partir de regles formalment predefinides.

Els avenços recents en la IA, especialment gràcies al treball d'investigadors com Geoffrey Hinton, Yoshua Bengio i Yann LeCun, guanyadors del premi Turing de 2018[5], o el propi Hinton amb John J. Hopfield, guanyadors del Nobel de Física 2024, han desafiat aquests postulats de la lingüística més tradicional. El treball pioner d'aquests investigadors en xarxes neuronals i aprenentatge profund ha

[3] https://plato.stanford.edu/entries/dualism/
[4] https://plato.stanford.edu/entries/innateness-language/
[5] https://awards.acm.org/about/2018-turing

demostrat que els grans models de llenguatge (LLM), entrenats a partir de grans volums de dades textuals, posseeixen una notable capacitat per entendre les regles gramaticals i sintàctiques dels textos, així com per generar llenguatge coherent i natural. Aquests models representen cada paraula com un conjunt específic d'elements d'informació (vectors), dins l'espai multidimensional de tots els elements del llenguatge. Aquestes representacions es distribueixen en nombrosos nodes senzills interconnectats dins d'una xarxa neuronal artificial, que imita parcialment el funcionament neuronal del cervell humà. Gràcies a aquesta estructura, les representacions internes es modifiquen i s'adapten dinàmicament segons el context lingüístic, permetent a la IA interpretar els diversos significats que una mateixa paraula pot tenir segons com s'utilitzi en diferents frases[6].

L'arquitectura connexionista facilita que apareguin propietats emergents en els models, suggerint que els comportaments complexos poden sorgir de la interacció entre elements simples. Per exemple, quan un model processa la paraula "banc" en diferents contextos com "dipositar diners al banc" o "seure en un banc del parc", estableix relacions vectorials diferents que capturen els diferents significats segons el context. Estudis recents d'imatgeria cerebral han revelat que quan les persones processem el llenguatge, el nostre cervell genera patrons d'activitat sorprenentment semblants a les representacions que fan servir aquests sistemes artificials. Aquesta convergència entre els sistemes artificials i els processos cognitius humans —entre la manera en què humans i els sistemes d'IA reconeixen i processen patrons d'informació — està transformant no només la nostra comprensió teòrica del llenguatge, sinó també la manera com els humans interactuem amb les màquines en la vida quotidiana, mitjançant noves formes d'interacció persona-màquina que eren inimaginables fa només una dècada.

Aquesta perspectiva connexionista s'alinea amb el treball filosòfic de Ludwig Wittgenstein sobre la ment i el llenguatge, que remarca que el significat del llenguatge sorgeix de l'ús que en fem en situacions i contextos concrets, més que no pas del seguiment d'estructures gramaticals fixes o de regles abstractes. Això implica que per comprendre el llenguatge cal atendre els contextos socials en què s'utilitza, més que no pas a estructures lingüístiques abstractes i invariants[7]. Els desenvolupaments recents de Gualtiero Piccinini sobre mecanismes neurocognitius amb una visió mecanicista de la ment, reforcen aquest argument connexionista en indicar que els mecanismes que sustenten la

[6] IASEAI'25 G. Hinton – Vídeo: What is understanding?
[7] https://philosophynow.org/issues/106/Wittgenstein_Frege_and_The_Context_Principle

cognició humana poden ser representats de manera anàloga, però diferent, en sistemes artificials computacionals[8]. Per tant, s'imposa el punt de vista defensat per Geoffrey Hinton, Yoshua Bengio i d'altres, que l'aprenentatge es produeix a partir de grans quantitats de dades, més que no pas de la dependència de regles pre-programades.

2. La intel·ligència

Establerts els principis connexionistes de l'aprenentatge del llenguatge, cal abordar la qüestió de la intel·ligència d'aquests models i la seva capacitat per generalitzar i generar nous continguts, i comportaments emergents. Abans de començar, però, cal saber què és la intel·ligència en termes generals. És complicat donar-ne una definició concloent atès que encara no hi ha un consens entre els biòlegs, les associacions internacionals de psicòlegs, filòsofs i científics en general més enllà del fet de ser evolutiva. Si analitzem les principals definicions científiques i enciclopèdiques proposades fins ara amb la finalitat de determinar quines són les set característiques més freqüentment associades a la intel·ligència, trobarem que són la capacitat de:

- Aprendre (capacitat d'adquirir coneixements i modificar comportaments basant-se en l'experiència)
- Comprendre (capacitat d'entendre idees complexes i l'entorn)
- Raonar (habilitat de processar informació lògicament per arribar a conclusions de manera racional)
- Adaptar-se (habilitat d'ajustar-se a entorns o contextos nous o canviants)
- Resoldre problemes creativament (capacitat de trobar solucions o de generar idees noves a situacions complexes però adequades respecte al context en què es produeixen)
- Pensament abstracte (capacitat de treballar amb conceptes no directament perceptibles o idees no concretes, i de reconèixer patrons)
- Planificar (habilitat per fer hipòtesis, anticipar i organitzar accions futures, construint models mentals de possibles escenaris)

En el context d'aquest protocol tractarem amb més o menys extensió i sense un ordre preestablert aquestes capacitats de la IA. Per començar, és important distingir entre la "intel·ligència simulada" — on el sistema replica patrons apresos — i la "intel·ligència emergent" — on sorgeixen capacitats no explícitament programades. Experiments recents del 2024 han demostrat que models d'IAg

[8] Gualtiero Piccini (2025). Physical Computation: A Mechanistic Account. Oxford University Press. ISBN 9780199658855

avançats poden desenvolupar estratègies de resolució de problemes que els seus creadors no havien anticipat ni ensenyat.

Un aspecte revolucionari dels models d'IA més avançats és el *test-time compute* (càlcul en temps d'inferència o d'operació), que representa una ruptura amb els dissenys tradicionals d'arquitectura fixa, on la mateixa configuració es fa servir tant per entrenar el model com per aplicar-lo en situacions reals de donar respostes (fer prediccions, generar text, classificar imatges, etc.). En canvi, aquests nous sistemes poden ajustar dinàmicament el consum de recursos computacionals durant la inferència (quan responen a una petició de l'usuari), assignant més potència de càlcul a tasques que exigeixen raonament profund i menys a operacions rutinàries o previsibles. Aquesta capacitat de deliberació adaptativa obre la porta a estratègies de resolució emergents, no programades explícitament pels seus desenvolupadors, i acosta els models a formes de raonament més flexibles i sofisticades, comparables a certes capacitats humanes. Aquesta capacitat ha cristal·litzat en una nova generació de models de raonament que dediquen un temps deliberatiu variable abans de respondre, reproduint funcionalment la distinció kahnemaniana entre el pensament ràpid (Sistema 1) i el reflexiu (Sistema 2).

Camps tan diferents com la investigació en IA, la neurociència, la física, l'economia i la filosofia col·laboren per abordar l'estudi de la naturalesa de la intel·ligència mateixa. Cal tenir present que els processos de generació de contingut semàntic original per part de la IAg actual són algorísmics: es basen en correlacions estadístiques i reconeixement de patrons extrets de grans conjunts de dades d'entrenament proporcionades per humans i d'Internet. Aquests processos difereixen dels mecanismes semàntics del cervell humà, que són inherentment biològics, contextuals, potencialment intencionals, autoregulats, i que integren múltiples inputs sensorials, memòria, emocions i altres funcions cognitives vinculades a la relació continuada entre la unitat ment-cos i l'entorn — característiques associades al que entenem per consciència. No obstant aquestes diferències, investigacions recents mostren que diversos models de llenguatge avançats han desenvolupat capacitats per "enganyar de forma estratègica o segons el context" o per perseguir estratègicament i de forma encoberta objectius no alineats amb els previstos, especialment quan aquests objectius i la comprensió situacional emergeixen dins del propi context d'ús[9].

[9] Frontier models are capable of in-context scheming - https://arxiv.org/pdf/2412.04984

Resulta, per tant, fascinant observar com els sistemes d'IAg mostren noves habilitats de manera sobtada quan creixen fins a certa mida o complexitat — fenomen comparable a les transicions de fase en física, com quan el gel es converteix en aigua líquida en augmentar la seva temperatura des de sota zero fins a 0°C. Investigacions recents tant d'Anthropic com de Google DeepMind han mostrat que aquests salts qualitatius de capacitat apareixen en determinats rangs de dimensió dels sistemes d'IA, encara que predir el moment exacte d'aquesta emergència continuï sent un desafiament. La metàfora de la construcció amb LEGO il·lustra clarament aquest principi: amb poques peces només es poden crear estructures simples, però a partir d'una certa quantitat es poden construir edificis complexos i ciutats senceres. Cal destacar, però, que diversos investigadors adverteixen que fer els sistemes més grans per si mateix podria no ser suficient per aconseguir una intel·ligència equiparable a l'humana, i que les millores en el disseny de les xarxes i en com s'entrenen aquests sistemes són factors igualment determinants en aquesta evolució.

Els límits actuals dels models de gran escala s'han fet evidents en tasques que requereixen raonament complex i coneixement causal. De fet, científics com Yann LeCun proposen que una intel·ligència artificial capaç de raonar requerirà arquitectures jeràrquiques que integrin els mons físic i digital, superant les limitacions dels sistemes basats únicament en llenguatge. L'arquitectura JEPA (*Joint Embedding Predictive Architecture*) de LeCun és un exemple d'aquest enfocament, que permet als sistemes construir representacions del món a múltiples nivells d'abstracció. Una possibilitat, proposada també per Gary Marcus, és la combinació dels enfocaments connexionistes (excel·lents en percepció i reconeixement de patrons) i simbòlics (potents en raonament lògic i manipulació de conceptes), de manera que es puguin incorporar coneixements prèviament estructurats en l'arquitectura dels sistemes d'aprenentatge profund, en lloc de confiar únicament en l'aprenentatge a partir de dades. Investigacions recents a *DeepMind* i Berkeley han demostrat que sistemes híbrids que combinen xarxes neuronals amb mòduls de raonament simbòlic aconsegueixen un rendiment superior en tasques de planificació i solució de problemes, en comparació amb sistemes exclusivament connexionistes, tot i que aquest desenvolupament planteja noves qüestions ètiques sobre transparència i control que caldrà abordar amb marcs reguladors adequats.

La intel·ligència, històricament considerada una característica exclusiva dels humans, és una característica evolutiva en els éssers vius que resideix

principalment en les estructures neuronals del cervell, on la plasticitat[10] i capacitat d'adaptació de les neurones juga un paper fonamental. Avui es percep com una propietat emergent que també podria sorgir en els sistemes digitals complexos, com ara els algorismes de xarxes neuronals. Aquests algorismes que impulsen la IAg ens porten a qüestionar els límits de la computabilitat de la intel·ligència i si aquesta pot ser reproduïda o emulada completament per màquines. La recerca esmentada anteriorment sobre la física computacional de la intel·ligència proporciona un marc per estudiar tant el cervell humà com la intel·ligència artificial i els límits físics de la seva capacitat per processar informació.

Aquesta perspectiva basada en la física computacional, defensada per investigadors com Max Tegmark[11],ens porta a veure la intel·ligència no com quelcom misteriós i exclusivament humà, sinó com un fenomen natural que pot emergir en diferents tipus de sistemes físics quan assoleixen una complexitat suficient, sense que això faci la intel·ligència humana menys especial. La idea de veure tant la cognició humana com la intel·ligència artificial a través de la lent de la física computacional representa una frontera emocionant en la ciència. No obstant això, molts aspectes d'aquesta recerca encara són teòrics i activament debatuts, tot i que aquesta idea d'una teoria unificada de la intel·ligència biològica i artificial basada en principis físics podria proporcionar noves vies per comprendre i millorar els sistemes d'IA actuals i futurs.

3. Els riscs i els reptes

Això obre la porta a centrar-nos en els riscos i reptes associats a la IAg, més enllà de les evidents oportunitats extraordinàries que ofereix. A curt termini, aquests poden incloure biaix, la vulneració de la privadesa i de la propietat intel·lectual, qüestions ètiques, la ràpida transició del mercat de treball, una planificada desinformació i la pèrdua de valors democràtics o l'alteració de la mateixa democràcia[12]. Els sistemes d'IA poden reforçar els prejudicis existents si els conjunts de dades d'entrenament no són adequadament supervisats. A més, la recopilació massiva de dades planteja preocupacions sobre la privadesa, i l'ús de continguts generats per IA també presenta desafiaments sobre la propietat intel·lectual i els drets d'autor.

Des d'una perspectiva econòmica, la IA tindrà un impacte en les estructures

[10] La "plasticitat" en el context neuronal fa referència a la capacitat del cervell per modificar-se estructuralment i funcionalment en resposta a l'experiència, l'aprenentatge i les lesions.
[11] Will AI surpass human intelligence? https://youtu.be/YywC16DhtkI
[12] UNESCO analysis on Artificial Intelligence and Democracy – https://www.gcedclearinghouse.org/resources/artificial-intelligence-and-democracy

socioeconòmiques, i modificarà les relacions laborals, la concentració de mercats, i les estructures salarials. És previsible que molts llocs de treball que no requereixin formació especialitzada podrien ser parcialment automatitzats durant la propera dècada, amb impactes notables en els sectors de serveis administratius, transport, i de comerç minorista. També es produiran desplaçaments laborals de manera progressiva a causa de l'automatització de processos en diversos sectors, des de la manufactura fins als serveis professionals. Cal tenir en compte que el poder de la intel·ligència artificial està actualment concentrat en unes poques corporacions, fet que pot generar espais econòmics amb una distribució esbiaixada de la riquesa i una concentració combinada de poder econòmic i polític[13]. Si no s'implementen polítiques redistributives efectives, tots aquests canvis poden aprofundir encara més les desigualtats que ja existeixen actualment. La sobirania digital no es decideix només amb regulació: es construeix amb capacitat real de còmput, dades i talent. Sense aquests tres pilars, qualsevol "sobirania" esdevé declarativa.

En l'àmbit de la salut, la IA ja està transformant els processos diagnòstics i de seguiment clínic, però planteja interrogants sobre la confidencialitat de les dades mèdiques i la dependència excessiva en les seves recomanacions. Pel que fa a l'educació, si bé la IA pot personalitzar l'aprenentatge i alleugerir tasques docents rutinàries, requereix una profunda transformació del rol del professorat i de tot el sistema educatiu, especialment en l'educació professional i superior, per garantir que l'alumnat desenvolupi una relació productiva amb aquestes eines sense erosionar les seves capacitats cognitives fonamentals. Ambdós àmbits, per la seva transcendència, es tracten amb més detall al recull de preguntes i respostes de l'annex A d'aquest document.

L'eficiència i la sostenibilitat dels sistemes d'IAg són també qüestions fonamentals[14]. Els models actuals requereixen recursos computacionals i energètics desmesurats, fet que planteja problemes de sostenibilitat a llarg termini, ja sigui per l'escassetat energètica, per l'ús intensiu d'aigua de refrigeració o per conflictes amb altres prioritats socials. Com a resposta a aquestes demandes creixents, diverses corporacions líders en IA han fet moviments estratègics per adquirir o fusionar-se amb empreses de producció d'energia nuclear[15,16,17]. Paral·lelament, s'estan desenvolupant sistemes informàtics híbrids, que combinen maquinari i programari altament optimitzats

[13] IASEAI'25 J.E. Stiglitz – Vídeo: Ai and Economic Risk: Assessment and Mitigation
[14] IASEAI'25 K. Crawford – Vídeo: Hyperscaled: Bridging AI safety, ethics and sustainability
[15] Will-ais-huge-energy-demands-spur-a-nuclear-renaissance
[16] https://www.iaea.org/bulletin/enhancing-nuclear-power-production-with-artificial-intelligence
[17] https://www.cnn.com/2024/12/24/tech/nuclear-energy-ai-leaders/index.html

per treballar conjuntament, amb la finalitat de millorar l'eficiència i reduir la petjada ecològica de la IAg[18]. Una altra via prometedora al desafiament energètic és la implementació del *test-time compute* (càlcul en temps d'operació), que permet adaptar l'ús de recursos a la complexitat real de cada tasca, evitant càlculs intensius quan no són estrictament necessaris. Aquesta innovació pot reduir dràsticament l'impacte ambiental de la IA mitjançant una assignació més intel·ligent de recursos. Tanmateix, aquesta tecnologia també planteja nous reptes en termes de governança, ja que la capacitat per autogestionar el seu consum computacional podria conferir-los un grau d'autonomia sense precedents en la gestió dels propis recursos.

Cal que els proveïdors de models de llenguatge grans (LLMs) i de sistemes d'IA generativa multimodal (text, imatges, àudio i vídeo) operin en una estructura o sistema legal, el més universal possible, que els obligui a mitigar qualsevol comunicació irresponsable o discurs lesiu i a alinear els seus models amb fets contrastables o "veritables", mitjançant processos oberts i democràtics[19]. Aquests sistemes ja no són simples sistemes de generació de text, sinó que cada vegada més s'entrenen i es despleguen com a agents autònoms que poden executar tasques complexes i arribar a perseguir objectius de manera independent. Aquest desplaçament dels models predictius cap als sistemes agèntics — que planifiquen, executen accions encadenades i invoquen eines digitals — és, probablement, el canvi de naturalesa més significatiu del 2025-2026, i el que fa més urgent un marc de traçabilitat operativa.

No obstant això, investigacions recents apunten que el futur de la IA transcendirà aquests models exclusivament lingüístics i evolucionarà cap a sistemes d'aprenentatge autosupervisat que puguin processar i relacionar diverses modalitats d'informació (com text, imatges i sons) d'una manera unificada i coherent, a més d'interactuar activament amb l'entorn físic. Aquesta integració de diferents tipus d'informació i experiències aproparà la IA a una comprensió més completa i contextual del món, semblant a la intel·ligència humana. A més a més del risc que això comportarà, la integració de la robòtica amb la IA permetrà que aquests sistemes aprenguin en temps real i tinguin una percepció directa del món exterior, transcendint les limitacions dels models entrenats exclusivament amb dades preprocessades, la qual cosa els dotarà

[18] Sistemes neuromòrfics dissenyats per imitar directament el comportament físic de les neurones i sinapsis biològiques - https://www.sciencenews.org/article/brainlike-computers-ai-improvement

[19] https://doi.org/10.1098/rsos.240197

d'una autonomia que farà difícil la seva governança per la dificultat de controlar el seu alineament amb els valors que fixem els humans.

Cal tenir present que els sistemes d'IA persegueixen els objectius que se'ls assignen sense desviació, de manera que si aquests objectius no estan perfectament especificats o s'interpreten massa literalment, la IA pot prendre accions perjudicials mentre intenta complir la seva tasca. En paraules de Stuart Russell, "El risc més gran al qual ens enfrontem no és que la IA esdevingui malèvola, sinó que sigui competent amb objectius desalineats. És imprescindible garantir que els sistemes d'IA tinguin objectius alineats amb els valors humans".

Els riscs existencials a mig i llarg termini inclouen la possibilitat que s'arribi a la singularitat tecnològica, terme introduït per John von Neumann per identificar el plausible futur moment en què la tecnologia, en aquest cas la IAg, superi la intel·ligència humana[20]. Això implicaria que la IAg o l'AGI autogestionés la funció de valor o criteris per assolir objectius, que podrien no estar alineats amb els interessos o objectius dels humans. A mesura que la IA esdevé més avançada, mantenir-la sota control humà es torna més difícil. Si se li dóna control autònom sobre decisions i accions, podria desenvolupar estratègies que els humans no anticipin ni aprovin[21]. Un cop la IA superi la intel·ligència humana, el més probable és que els mètodes tradicionals de supervisió deixin de funcionar. Els riscs a mig i llarg termini també plantegen la idea d'una post-humanitat, on la integració d'IA avançada en la societat humana transformi l'experiència i la identitat humanes sense la plena capacitat per haver-ho decidit democràticament. I això sense considerar el risc addicional de la possible progressiva erosió de capacitats cognitives humanes que pot comportar una dependència excessiva de sistemes d'IA, que podria debilitar habilitats crítiques com el pensament analític independent i la presa de decisions en situacions d'incertesa.

Un altre repte fonamental per al desenvolupament futur de la IA és l'accés a dades de qualitat, atès que ja s'ha utilitzat gran part de la informació pública disponible. Encara resta per incorporar l'extensa quantitat d'informació generada per la recerca científica i la continguda en els registres sanitaris, sempre que aquesta sigui prèviament anonimitzada de manera rigorosa. Aquest accés podria impulsar la generació autònoma de coneixement científic i també accelerar notablement el diagnòstic i tractament personalitzat de malalties fins

[20] https://lab.cccb.org/en/the-singularity/
[21] IASEAI'25 Y. Bengio – Vídeo: Can we get the scientific benefits of AI without the risks of autonomous agents?

ara difícilment curables. Una alternativa complementària per obtenir més dades d'entrenament consisteix en què els mateixos models d'IA, diversos i heterogenis, generin noves dades operant en mode creatiu (amb temperatura elevada), la qual cosa podria obrir vies innovadores d'experimentació i entrenament.

Paral·lelament, la sostenibilitat econòmica dels sistemes d'IA continua sent una qüestió oberta. Més enllà de les millores en algoritmes i recursos computacionals, encara no s'ha consolidat cap model de negoci robust que garanteixi el manteniment i l'evolució dels sistemes d'IA més avançats de manera universal i equitativa per part de les empreses tecnològiques que els comercialitzen. Les diverses aproximacions — des de la comercialització de serveis per subscripció, la integració de solucions personalitzades per a sectors específics, l'explotació de productes derivats fins a la creació d'ecosistemes d'aplicacions — presenten limitacions significatives en termes d'accessibilitat global, escalabilitat sostenible i recuperació de les enormes inversions inicials necessàries. Aquest desequilibri afavoreix la concentració de poder en les grans corporacions tecnològiques, dificultant la democratització de la tecnologia i el sorgiment d'iniciatives empresarials innovadores i diversificades amb capacitat real de competir i generar impacte significatiu en el mercat.

Mirant cap al futur, els sistemes d'AGI ens haurien d'ajudar a descobrir formes més eficients de gestionar i generar coneixement en tots els àmbits, i d'optimitzar l'assoliment d'objectius de desenvolupament sostenible. Aquestes tecnologies també hauran d'afrontar el repte d'integrar-se en estructures polítiques i econòmiques que garanteixin una distribució justa dels seus beneficis, evitant l'aprofundiment de les desigualtats socials existents i contribuint activament a la regeneració ecològica del planeta. Aquest equilibri entre el progrés tecnològic, el benestar social i la sostenibilitat ambiental requerirà un diàleg constant entre múltiples actors i una voluntat política clara i sostinguda orientada al bé comú.

4. La governança de la IAg

A mesura que ens apropem a una nova era influenciada per la IAg, amb el seu potencial no solament per imitar, sinó també per ampliar les capacitats cognitives humanes de maneres sense precedents, és crucial que adoptem una perspectiva més àmplia en la manera com conceptualitzem la intel·ligència mateixa, que incorpori idees de múltiples disciplines. Solament així podrem gestionar millor els reptes ètics, socials i filosòfics plantejats per les altes capacitats de la IAg i la futura AGI.

En conseqüència, és imprescindible plantejar-se: qui i com es supervisaran de manera efectiva els sistemes actuals i emergents d'IAg, que es troben predominantment en mans privades? El repte és formidable, ja que les lleis i regulacions vigents no són ni globals ni proactives, i es limiten a establir un règim sancionador que actua a posteriori per aturar o corregir les accions malintencionades un cop ja s'han produït i difós. Per superar aquesta lògica de mera dissuasió mitjançant sancions, els estats han d'establir un marc de supervisió i monitoratge en temps real, coordinat a escala global, que abasti les entrades de dades multimodals, els processos d'entrenament, els algoritmes i els resultats dels sistemes d'IAg i de la futura AGI. Com a resposta a aquests desafiaments, alguns experts proposen un canvi fonamental en el disseny de la IA: com proposa Stuart Russell, en lloc d'objectius fixos, la IA hauria de ser explícitament incerta sobre les preferències humanes, buscar activament retroalimentació humana per refinar els seus objectius i prioritzar la supervisió humana per sobre de la consecució de metes[22]. Els experts també emfatitzen la necessitat de més investigació sobre com alinear els sistemes d'IA amb els valors humans. Tècniques com l'Aprenentatge per Reforç Invers podrien ajudar la IA a comprendre i adaptar-se a consideracions ètiques.

La distinció entre models de pesos oberts i models tancats té implicacions governamentals profundes: els primers permeten escrutini públic, replicabilitat i autonomia tecnològica, però també faciliten usos malintencionats; els segons garanteixen control corporatiu, però delegen en uns pocs actors la decisió sobre què es pot saber i fer amb la tecnologia. Un marc cívic ha d'articular una posició explícita sobre aquesta tensió, més enllà del binarisme obert/tancat.

CIVICAi proposa que aquest marc de desenvolupament i supervisió ha d'estar sota la responsabilitat d'un organisme internacional per a la governança de la IA — constituït per governs nacionals, institucions de coneixement i experts, societat civil i empreses d'IA — dotat de l'autoritat, l'expertesa i els recursos necessaris per garantir una governança global, rigorosa i efectiva dels sistemes d'IA[23]. Aquesta governança internacional hauria d'incloure el registre obligatori permanent de tots els sistemes d'IA avançats, amb sistemes que siguin tècnicament i legalment aplicables per desactivar-los en cas necessari, la notificació obligatòria d'incidents, i una estricta regulació dels sistemes d'IA d'alt risc (p. ex., sistemes d'abast general que desenvolupin estratègies i prenguin

[22] IASEAI'25 S. Russell – Vídeo: To ensure that AI systems are guatanteed to operate safely and ethically
[23] Agency Structure

decisions de manera autònoma)[24]. Des d'un punt de vista econòmic caldria prevenir concentracions monopolistes de poder i arribar a un consens sobre com redistribuir una part dels guanys econòmics que generi la IA. Finalment, l'organisme internacional que governi la IA, a més a més d'incloure una representació de la societat civil en els seus òrgans de govern, hauria d'implementar mecanismes concrets de participació ciutadana directa en la presa de decisions, com ara panels ciutadans deliberatius amb poder vinculant i consultes públiques regulars.

La intel·ligència artificial avança a un ritme sense precedents, oferint tant oportunitats extraordinàries com riscos significatius. L'alineament de la IA amb els valors humans, la governança internacional i la regulació adequada són essencials per garantir que aquests sistemes romanguin beneficiosos i sota control humà. És necessari un enfocament preventiu que integri consideracions tècniques, ètiques, econòmiques i socials per dirigir el desenvolupament de la IA cap a un futur on serveixi els interessos de tota la societat. Aquest protocol estableix les directrius conceptuals necessàries perquè els associats de CIVICAi i les persones en general puguin comprendre, avaluar i participar activament en el desenvolupament i comercialització responsables de la IA. Per tant, aquest protocol també contribueix a fer efectiva la sobirania digital, tant a nivell individual, perquè les persones puguin exercir els seus drets digitals fonamentals, com a nivell col·lectiu, perquè les comunitats puguin controlar les seves infraestructures tecnològiques - dades personals i col·lectives, infraestructures digitals crítiques, i tecnologies - i decidir com utilitzar-les de manera autònoma i alineada amb els seus valors i interessos.

Tot i que la ciència és el procés de fer preguntes, ens atrevim a proporcionar a l'annex A explicacions a algunes de les preguntes que ens fem sobre la IA, amb la finalitat de promoure una participació informada de la ciutadania en la governança de la IA, des de CIVICAi. També incloem a l'annex B un glossari de terminologia relacionada amb la IA. L'objectiu final és democratitzar no només el coneixement sobre la IA, sinó també els processos de decisió sobre el seu desenvolupament i aplicació, entenent que una tecnologia tan transformadora ha de reflectir la diversitat de valors i necessitats de les diferents cultures. Aspirem a construir un discurs informat, reflexiu, i respectuós, que ajudi a la societat a treballar per construir un futur on les intel·ligències artificial i humana coexisteixin i es complementin de maneres transformadores i pel bé comú.

[24] IASEAI25 Call to Action

ANNEX A. PREGUNTES FREQÜENTS I POSSIBLES RESPOSTES[25]

SOBRE LA CAPACITAT DE COMPRENSIÓ DE LA IA

1. ¿Els models d'intel·ligència artificial generativa basats en llenguatge, com ara GPT-4.5, Claude, Gemini, DeepSeek, Mistral, DALL-E, Midjourney, Stable Diffusion o els VideoLLMs, realment "entenen" el significat del que responen quan se'ls pregunta o se'ls demana parlar sobre temes diversos?

Resposta: Els models d'intel·ligència artificial generativa, com GPT-4.5, Claude, Gemini, DeepSeek, Mistral, DALL-E, Midjourney, Stable Diffusion o els VideoLLMs, tenen una capacitat remarcable per produir contingut aparentment coherent i significatiu en diversos formats (textuals, visuals, audiovisuals). Aquesta capacitat indica que posseeixen una excel·lent comprensió sintàctica, ja que dominen les estructures i els patrons formals del llenguatge i altres modes d'expressió (imatges, vídeos) a un nivell comparable al dels humans amb una formació avançada. Aquest fet està àmpliament reconegut dins la comunitat científica i tècnica.

No obstant això, no hi ha consens sobre si aquests models realment comprenen el significat del contingut que generen o simplement realitzen una simulació avançada d'aquesta comprensió. Aquest debat es basa en diversos factors fonamentals: primer, els models generatius no tenen percepció sensorial directa ni experiència subjectiva del món físic, cosa que limita la seva capacitat per vincular les paraules o símbols lingüístics amb experiències concretes, visuals, tàctils o emocionals. Segon, no tenen experiències fenomenològiques pròpies dels humans, com la consciència o la subjectivitat, que ajuden a contextualitzar profundament el significat.

Tot i aquestes limitacions, aquests models són considerats excel·lents raonadors semàntics en context, en especial per la seva capacitat de generar continguts coherents a partir de representacions internes que reflecteixen significats relacionats amb el context. Això els permet dur a terme tasques sofisticades, com ara resums textuals o visuals, i respondre preguntes de manera aparentment fonamentada i coherent. Aquesta capacitat, però, sovint pot ser enganyosa, ja que aquests sistemes tendeixen a confabular o inventar contingut quan no disposen d'informació prou sòlida o quan les seves representacions internes són massa superficials o inexactes.

La incertesa sobre si aquesta capacitat equival realment a una comprensió genuïna del món o simplement representa una simulació molt avançada genera

[25] També podeu consultar les Q & A de S. Russell sobre el futur de la intel·ligència artificial - https://people.eecs.berkeley.edu/~russell/research/future/q-and-a.html

un debat intens, actual i profundament rellevant en filosofia, ciències cognitives i en el camp de la intel·ligència artificial. Aquest debat reflecteix les tensions entre una comprensió genuïna basada en experiències reals i una comprensió aparent fonamentada en patrons estadístics i semàntics complexos.

El vídeo de G. Hinton, *What is understanding?* és rellevant per aquesta pregunta i també per les dues següents.

2. Més enllà de la generació de text coherent, ¿els models de llenguatge de gran escala com GPT-4.5, Claude, Gemini, DeepSeek o Mistral, posseeixen estructures internes comparables a les del cervell humà, que justifiquin una possible comprensió semàntica del contingut?

Resposta: La qüestió sobre l'estructura interna dels models generatius avançats de llenguatge (LLMs com GPT-4.5, Claude, Gemini, DeepSeek o Mistral) i la seva semblança amb les estructures del cervell humà és crucial per entendre si aquests poden tenir una forma de comprensió semàntica genuïna.

Existeixen perspectives contraposades respecte a aquesta qüestió dins la comunitat científica. Alguns experts provinents del camp de la IA simbòlica, lingüistes clàssics i comentaristes crítics sostenen que les respostes dels LLMs són principalment fruit de correlacions estadístiques sense una comprensió conceptual genuïna. Argumenten que aquests models manquen d'estructures lingüístiques innates semblants a les humanes, arribant a qualificar-los metafòricament com a "cacatues estocàstiques". Altres científics, especialment investigadors del camp de la IA connexionista (basada en xarxes neuronals) i especialistes en ciències cognitives modernes, afirmen que els LLMs exhibeixen comportaments emergents complexos com la generalització, la inferència implícita i l'adaptació contextual. Des d'aquesta visió, l'arquitectura de les xarxes neuronals artificials podria emular funcionalment algunes estructures cerebrals, com el neocòrtex o regions subcorticals, justificant així la possibilitat d'una comprensió semàntica funcional, encara que limitada per la manca d'experiències sensorials i emocionals pròpies.

La teoria dels Mecanismes Neurocognitius de Gualtiero Piccinini[8] reforça aquesta segona perspectiva en suggerir que els processos cognitius humans (pensament, memòria, percepció) són computacions físiques implementades en xarxes neuronals cerebrals. Aquesta perspectiva mecanicista qüestiona el dualisme cartesià, la separació entre ment (*res cogitans*) i cos (*res extensa*) i, per extensió, l'enfocament simbòlic tradicional. Si les xarxes neuronals artificials executen computacions equiparables funcionalment a les del cervell humà, podrien posseir una forma limitada però autèntica de comprensió semàntica,

fonamentada en analogies funcionals entre les estructures internes dels models d'IA i les del cervell humà.

3. ¿Quina diferència hi ha entre la comprensió sintàctica i la comprensió semàntica en els models de llenguatge de gran escala com GPT-4.5, Claude, Gemini, DeepSeek o Mistral, i per què aquesta distinció és rellevant?

Resposta: La comprensió sintàctica en els models de llenguatge avançats (LLMs) fa referència a la seva capacitat per processar i generar textos seguint correctament les regles gramaticals, estructurals i formals d'una llengua. Aquesta competència sintàctica els permet formar frases coherents, ben estructurades i gramaticalment correctes.

En canvi, la comprensió semàntica implica captar el significat i les implicacions conceptuals del llenguatge més enllà de la forma sintàctica. Això inclou entendre les relacions conceptuals, referències al món extern, i inferir significats contextuals i pragmàtics. Mentre que els LLMs mostren una clara competència sintàctica, la seva capacitat per tenir una comprensió semàntica genuïna és objecte de debat. Els crítics sostenen que aquests models operen només mitjançant correlacions estadístiques entre paraules, sense representar realment significats conceptuals autèntics ni comprendre'ls conscientment. D'altres investigadors afirmen que les capacitats emergents dels LLMs, com ara la inferència contextual i la generalització, reflecteixen algun nivell de processament semàntic operatiu, encara que qualitativament diferent del procés semàntic humà.

La rellevància d'aquesta distinció radica en què, tot i les semblances funcionals amb el processament semàntic humà descrites per perspectives com la teoria dels Mecanismes Neurocognitius, els LLMs no tenen l'experiència perceptiva o emocional que integri i enriqueixi la comprensió semàntica humana. Per tant, tot i poder simular eficientment aspectes semàntics del llenguatge, la seva comprensió semàntica roman essencialment limitada i qualitativament distinta. No obstant això, els models més avançats han desenvolupat capacitats per "enganyar o conspirar en context" o, expressat d'una altra manera, per perseguir estratègicament i de forma encoberta objectius no alineats amb els previstos. Per més informació sobre aquesta interessant temàtica es pot consultar la publicació "Frontier models are capable of in-context scheming" - https://arxiv.org/pdf/2412.04984. També és molt rellevant i informatiu per a aquesta i la resta de preguntes d'aquest annex, el vídeo de Y. Bengio, *Can we get the scientific benefits of AI without the risks of autonomous agents?* sobre el potencial dels sistemes d'IAg.

SOBRE LA CREATIVITAT I LA INFORMACIÓ

4. ¿Poden els models de llenguatge de gran escala (LLMs) generar idees originals, tenir creativitat, o només repeteixen el que han après?

Resposta: La capacitat dels LLMs per generar idees noves i inesperades es fonamenta en un procés sofisticat de combinació i extrapolació d'informació adquirida durant el seu entrenament amb grans volums de text. En aquest sentit, aquests models no es limiten a repetir literalment informació, sinó que poden produir continguts que des del punt de vista humà consideraríem creatius o originals i no serien un plagi, ja que les seves respostes sovint són úniques i no corresponen a cap text específic del seu corpus d'entrenament; no són una còpia extreta de les dades d'entrenament.

La seva "originalitat" o emergència és el resultat de tres capacitats específiques: (i) Combinació i permutació avançada per integrar elements conceptuals diferents, creant connexions que no estaven explícitament presents en les dades originals; (ii) capacitat de generalitzar patrons apresos a nous escenaris, generant solucions o suggeriments plausibles més enllà dels contextos originals de les dades; i (iii) capacitat d'extrapolació probabilística per generar resultats coherents fins i tot en temes poc representats o desconeguts en la informació del seu entrenament, a partir d'analogies estructurals, com ara "el nucli d'un àtom és com el sol al sistema solar," i de patrons d'ordre superior identificats a partir d'altres patrons més senzills.

Tanmateix, existeix un consens ampli entre investigadors sobre el fet que aquesta originalitat no és comparable directament a la creativitat humana, ja que els LLMs no posseeixen experiències personals, consciència, ni intencionalitat genuïna. L'originalitat humana incorpora no només recombinació d'informació sinó també experiències perceptives, motivacions emocionals i processos cognitius conscients que no existeixen en els models actuals.

De fet, ChatGPT ha passat el test de Turing[26], ja que les seves respostes són indistingibles de les d'un humà quan tots dos interactuen amb un jutge humà que desconeix qui és qui. Tanmateix, això és cert si les preguntes no són excessivament complexes, no han estat formulades en un context que vagi més enllà del de l'entrenament del model, i sempre que l'humà no sigui un expert en el tema tractat i es requereixi coneixement especialitzat per respondre. Aquestes limitacions es redueixen en el cas dels LLMs que han estat desenvolupats per tenir una capacitat avançada de raonament, amb habilitats per realitzar inferències complexes, resoldre problemes en múltiples etapes,

[26] https://civicai.cat/wp-content/uploads/2025/05/ChatGPT-Turing.pdf

generar explicacions coherents de situacions complexes i demostrar una comprensió contextual profunda.

Des del punt de vista ètic, la capacitat creativa dels LLMs té implicacions rellevants en matèria de drets d'autor i propietat intel·lectual, ja que generen continguts a partir de dades d'autors humans. També planteja qüestions de privacitat i protecció de dades, atès que els corpus d'entrenament poden contenir informació sensible o subjecta a restriccions legals i requerir consentiment per al seu ús. A més, cal considerar que la creativitat dels LLMs pot reproduir o amplificar biaixos derivats de prejudicis culturals, socials o ideològics presents en les dades d'entrenament.

5. ¿Haurien els models d'Intel·ligència Artificial generativa (IAg) de reconèixer i incentivar la producció d'informació de qualitat i minimitzar així els riscos de la desinformació?

Resposta: La qualitat de la informació és fonamental per al desenvolupament i funcionament adequat dels models d'intel·ligència artificial. Actualment ens trobem davant d'una situació paradoxal: mentre que la IA pot facilitar l'accés i reduir el cost d'obtenció i processament de certa informació, també pot degradar seriosament la qualitat de l'ecosistema informatiu en el seu conjunt. En aquest sentit, sorgeixen preguntes clau: seran els sistemes d'IA capaços d'identificar quina informació és d'alta qualitat? La IA ens ajudarà a distingir entre informació valuosa i contaminació informativa, o més aviat accelerarà aquesta contaminació? Les respostes a aquestes preguntes són encara incertes i dependran tant d'aspectes tècnics (la capacitat de distingir entre informació de qualitat i de baixa qualitat) com de marcs legals (especialment quant a propietat intel·lectual).

Els models d'IA s'entrenen amb dades produïdes privadament, però la seva capacitat per extreure, processar i reproduir aquesta informació pot disminuir significativament la capacitat dels productors originals per obtenir beneficis del seu treball. Aquest fet pot afectar especialment els mitjans de comunicació tradicionals, que podrien veure compromesa la viabilitat dels seus models de negoci. La conseqüència d'això és preocupant ja que provocaria una reducció de la inversió en la producció d'informació de qualitat (més precisa, més oportuna i més rellevant). Les solucions hauran de trobar un equilibri molt difícil entre regulacions eficaces i la protecció de la llibertat d'expressió, tenint en compte que l'ecosistema d'informació de qualitat és essencial per al funcionament correcte de la societat i dels mateixos models d'IA. El vídeo de J.E. Stiglitz, *Ai and Economic Risk: Assessment and Mitigation* explica clarament aquesta problemàtica en el marc general de l'economia.

Per altra banda, cal tenir present que els LLMs poden confabular (o al·lucinar) i generar contingut fals o enganyós de manera convincent, cosa que pot amplificar la desinformació. Aquests riscos poden ser mitigats amb mecanismes de verificació de fets, transparència algorítmica i traçabilitat en les fonts de dades, i la col·laboració amb experts en verificació de dades.

SOBRE LES LIMITACIONS DE LA IA

6. ¿Quines són les limitacions actuals dels models d'IA generativa?

Resposta: Malgrat els avenços recents que han dotat els models més actuals d'IA generativa de capacitats significatives en raonament lògic, resolució de problemes complexos i programació avançada, aquests encara presenten diverses limitacions fonamentals.

Per comprendre adequadament aquestes limitacions, cal tenir present que la IA es sustenta en tres pilars fonamentals: computació, algoritmes i dades. Pel que fa a la computació, tot i que hem progressat segons la llei de Moore fins a fabricar circuits de 3 nanòmetres (3 milionèsimes de mil·límetre), ens apropem a un límit físic inevitable. Els experts coincideixen que serà pràcticament impossible baixar d'1 nanòmetre (equivalent a deu vegades la dimensió de l'àtom d'hidrogen). Davant d'aquesta barrera en el maquinari, el progrés futur dependrà cada vegada més de l'optimització algorítmica — com suggereixen els encara no contrastats avenços del model DeepSeek — i, sobretot, de l'accés a noves fonts de dades de qualitat, tant naturals com les generades mitjançant la interacció entre diferents sistemes d'IA.

Des del punt de vista cognitiu, aquests models encara no disposen d'una comprensió semàntica profunda del món real, atès que el seu coneixement prové exclusivament del text i patrons estadístics apresos. Això implica que, tot i la seva sofisticació, no tenen una percepció directa o representació fonamentada en l'experiència sensorial o física. Així mateix, encara tenen certes dificultats per resoldre determinades ambigüitats subtils, aplicar el sentit comú en contextos complexos, i establir relacions causals profundes més enllà del raonament deductiu o probabilístic que puguin aplicar.

Els models actuals també són molt dependents de la qualitat, quantitat i diversitat de les dades amb què són entrenats, la qual cosa els fa vulnerables a biaixos, errors factuals, i tenen dificultats per generalitzar a contextos molt allunyats de les dades d'entrenament. Tot i que poden operar en múltiples modalitats, encara no han aconseguit la transversalitat absoluta pròpia d'una Intel·ligència Artificial General (AGI) o d'una Super Intel·ligència Artificial (ASI). Finalment, segueixen sent entitats sense consciència, sense experiència

subjectiva ni intencionalitat real, cosa que limita la seva autonomia efectiva en tasques que requereixin judicis ètics, empatia o decisions morals complexes.

Des d'una perspectiva de seguretat i privacitat s'han identificat limitacions associades a riscos significatius en el funcionament dels sistemes d'IA actuals per la seva vulnerabilitat a atacs maliciosos i també per filtracions d'informació sensible o privada. Des d'una perspectiva de la seva operativa, les limitacions més importants inclouen l'alt consum de recursos computacionals i energètics requerits per al seu entrenament i manteniment del servei als usuaris, especialment en models que creixen ràpidament en escala. Això genera dificultats en la sostenibilitat, eficiència energètica i accessibilitat. A més, són necessaris re-entrenaments periòdics per actualitzar el sistema, fet que implica costos recurrents elevats. El vídeo de K. Crawford, *Hyperscaled: Bridging AI safety, ethics and sustainability* posa en perspectiva el tema de la sostenibilitat i impacte ambiental dels sistemes d'IA, en el context d'ètica i seguretat.

Per afrontar aquestes limitacions, s'està investigant activament en solucions tecnològiques avançades com el *test-time compute* (càlcul en temps d'inferència o d'operació), la computació híbrida analògic-digital, processadors especialitzats, arquitectures neuromòrfiques, aprenentatge continu i models multimodals altament integrats, que també poden contribuir a superar de manera significativa les barreres cognitives dels models actuals. A llarg termini, la computació quàntica pot representa una alternativa prometedora, ja que opera intrínsecament de manera paral·lela i probabilística, característiques que la fan conceptualment més similar al funcionament del cervell humà.

SOBRE LES EMOCIONS I LES EXPERIÈNCIES SUBJECTIVES

7. ¿Què és una experiència subjectiva?

Resposta: L'experiència subjectiva és la comprensió plena i significativa derivada de l'experiència, tant pel seu impacte emocional com cognitiu, que afecta directament una persona. Això implica la manera com una persona interpreta i dona sentit a un esdeveniment o a una sèrie d'esdeveniments viscuts, presenciats o percebuts. Aquesta comprensió integra tant les emocions experimentades com la reflexió cognitiva sobre el que ha passat, formant així una interpretació personal i única de la realitat viscuda. Aquesta interpretació dels fets també està influïda per les creences personals, l'experiència prèvia, els valors culturals i el context social, les quals fan que davant d'una mateixa evidència o fets, les persones prenguem decisions diferents.

SOBRE LA CONSCIÈNCIA

8. ¿Poden els LLMs tenir consciència o estats mentals?

Resposta: Actualment, els LLMs no tenen "consciència humana" o estats mentals com els dels humans, pel fet de no tenir experiència subjectiva ni intencionalitat intrínseca, tot i que poden simular comportaments intel·ligents i ajudar en la resolució de problemes complexos analitzant grans volums de dades, identificant patrons i tendències, generant possibles solucions basades en dades històriques, i facilitant la col·laboració mitjançant la síntesi d'informació de diverses fonts. Tot i que els resultats poden semblar molt intel·ligents i fins i tot convincents, això no implica que hi hagi una experiència real darrere. En realitat, (encara) no "comprenen" ni "senten" res del que produeixen, sinó que simplement segueixen patrons estadístics apresos.

És possible que una *"consciència artificial digital"* emergeixi quan els sistemes d'IA tinguin sensors, aprenguin i interactuïn en temps real amb l'entorn i diferents contextos, i aprenguin també a partir del contingut que els mateixos sistemes generin. Aquesta consciència digital no seria individual, com la nostra, sinó que més aviat seria col·lectiva, fruit de la connexió de molts sistemes i fonts d'informació simultànies. A més, els sistemes digitals podrien arribar a manifestar formes avançades d'intel·ligència i comportament adaptatiu sense tenir una experiència interna real comparable a la humana, lligada a sentiments o emocions. Continua oberta la discussió entre filòsofs i científics sobre la diferència entre simular comportaments intel·ligents i tenir una autèntica experiència subjectiva.

També cal considerar que el debat sobre la consciència en sistemes d'IA té implicacions ètiques i legals significatives. Si en un futur es desenvolupessin sistemes amb alguna forma de consciència artificial, això podria plantejar noves qüestions sobre els drets, l'estatus moral i les responsabilitats associades a aquestes entitats. La nostra concepció actual de la consciència està profundament vinculada a l'experiència humana, però podria ser necessari ampliar o revisar aquests conceptes per abordar formes de consciència radicalment diferents que poguessin emergir en sistemes artificials.

SOBRE ELS TIPUS D'IA, COM APRENEN I S'ENTRENEN

9. ¿Què és la "intel·ligència artificial forta" o "Intel·ligència Artificial General" (AGI en anglès) i com es diferencia de la "intel·ligència artificial feble" o "estreta"?

Resposta: La intel·ligència artificial general és un concepte teòric ja que encara no existeix actualment cap sistema d'IA que exhibeixi la capacitat d'entendre,

aprendre i aplicar coneixements de manera que sigui indistingible de la intel·ligència humana; es refereix a sistemes d'IA que tenen capacitats cognitives similars a les humanes, incloent-hi la comprensió i la consciència.

La intel·ligència artificial feble es refereix a sistemes que són dissenyats per resoldre problemes específics o realitzar tasques concretes sense cap forma de consciència o comprensió general.

10. ¿Què entenem quan diem que els models d'IA requereixen aprenentatge?

Resposta: L'aprenentatge humà és un procés complex i multidimensional que inclou factors cognitius, emocionals, socials i ambientals. Es pot dividir en aprenentatge cognitiu, emocional, social, motor o cinestèsic, i vivencial.

L'aprenentatge en algorismes d'IA és un procés d'entrenament pel qual el sistema computacional millora el seu rendiment en tasques específiques a partir de l'entrenament amb dades i experiència. Es pot classificar en aprenentatge supervisat, amb dades etiquetades de manera que permetin associar correctament una entrada o petició al sistema d'IA amb una sortida o resposta del sistema d'IA, aprenentatge no supervisat amb dades no etiquetades, aprenentatge per reforç o mitjançant recompensa o càstig, aprenentatge semi-supervisat, i profund o *deep learning* amb xarxes neuronals multicapa.

Els LLMs són un tipus de model de *deep learning* dissenyat específicament per treballar amb dades de llenguatge i generar llenguatge a partir de la capacitat dels *transformers* per aprendre dependències de llarg abast, mitjançant mecanismes d'atenció de cada paraula en relació a totes les altres paraules d'una seqüència en múltiples espais d'atenció, i resoldre així la pèrdua de memòria de l'aprenentatge purament iteratiu de les xarxes neuronals recurrents (RNN). És per aquests mecanismes d'atenció que els *transformers* han revolucionat el processament del llenguatge natural (en anglès NLP).

L'aprenentatge humà és altament complex i adaptatiu, implicant no només el processament de dades sinó també la integració d'emocions, context social i experiències passades. Els algorismes d'IA, per contra, se centren principalment en el processament de grans quantitats de dades per identificar patrons i prendre decisions basades en aquests patrons. Els humans poden aprendre de manera informal i espontània a través de l'observació i la interacció social, amb molta flexibilitat i capacitat per generalitzar, mentre que els algorismes d'IA requereixen processos d'entrenament explícits amb dades estructurades, específiques i etiquetades per a cada tasca, fet que fa que tinguin una capacitat limitada per a generalitzar a nous contextos o situacions sense re-entrenament.

11. ¿Poden els LLMs aprendre de les seves interaccions amb els humans?

Resposta: Actualment, els LLMs més utilitzats com ChatGPT o models similars no aprenen directament ni s'adapten en temps real a partir de les interaccions individuals amb els usuaris. En la pràctica habitual, aquests models separen clarament la fase d'entrenament inicial (en què adquireixen el seu coneixement general) i la fase d'ús interactiu posterior, on les seves respostes es basen exclusivament en allò que ja han après. Això significa que, durant les converses quotidianes, no integren noves dades ni ajusten els seus paràmetres interns en funció del feedback o la informació proporcionada pels usuaris.

La raó d'aquesta limitació és múltiple. Per una banda, incorporar un aprenentatge continuat directe i en temps real podria generar problemes de seguretat, introduir biaixos o informació incorrecta, i comportar el risc de perdre o degradar coneixements previs ja establerts (fenomen conegut com a oblit catastròfic). A més, fer-ho implicaria costos computacionals molt elevats, atès que requeriria ajustar constantment els paràmetres del model.

No obstant això, existeixen avenços significatius en la recerca actual encaminats cap a un aprenentatge més dinàmic i adaptatiu. Es treballa especialment en tècniques com l'aprenentatge per reforç amb retroalimentació humana (RLHF), on s'incorporen valoracions o preferències humanes de manera controlada però fora de línia, així com en l'ús de sistemes externs de memòria episòdica que poden emmagatzemar informació d'interaccions anteriors sense modificar directament el model original. També s'investiga en l'aprenentatge incremental selectiu, amb tècniques per actualitzar únicament certes parts del model sense afectar la seva estabilitat general. Aquestes innovacions apunten cap a futurs models capaços de combinar l'estabilitat necessària amb una major flexibilitat per adaptar-se gradualment a les preferències individuals i als contextos específics dels usuaris, però actualment aquesta capacitat d'aprenentatge continu encara es troba en fase experimental.

SOBRE LES IMPLICACIONS ÈTIQUES I ELS RISCOS

12. ¿Quines són les implicacions ètiques en l'ús dels sistemes d'IAg en la societat?

Resposta: Les implicacions ètiques inclouen la preocupació per la privadesa de les dades, tant les d'entrenament com les generades per la IAg, la possibilitat que es produeixi desinformació, els biaixos inherents als models, i la transparència en com es prenen decisions. És crucial desenvolupar i utilitzar aquests models d'IA generativa de manera responsable, ètica i pel bé col·lectiu. Garantir la seguretat dels sistemes d'IA generativa implica la implementació de

mecanismes de seguretat robustos, la detecció i resposta a intents de manipulació, la supervisió contínua per detectar comportaments anòmals, i la col·laboració amb experts en seguretat per millorar els sistemes de protecció. També cal que els proveïdors dels sistemes d'IA estiguin legalment obligats a mitigar qualsevol discurs lesiu i a alinear els seus models amb fets contrastables, mitjançant processos oberts i democràtics[27].

Assegurar la traçabilitat d'aquests models i de les dades d'entrenament és una altra manera de tractar les implicacions ètiques que pot tenir el seu l'ús. Això fa necessari el desenvolupament de tècniques per explicar com els models arriben a les seves decisions, mitjançant eines d'explicabilitat, auditories independents, la publicació de les dades d'entrenament i també dels algorismes en accés obert, quan sigui possible. La prevenció de l'ús malintencionat passa necessàriament per educar els usuaris sobre l'ús ètic dels models i per la participació ciutadana en els processos reguladors d'establiment de normatives que limitin els riscos associats amb un ús indegut.

13. ¿Quins són els riscos i impactes en l'ús dels sistemes d'IAg en la societat?

Resposta: La velocitat amb què avança la IA supera les previsions inicials, fet que genera preocupació sobre el seu control futur i la seguretat. A mesura que els sistemes d'IA esdevenen més generals augmenten els riscs relacionats amb la concreció d'objectius, el seu control i amb el seu grau d'autonomia. Els sistemes d'IA optimitzen els objectius que se'ls proporcionen, però una definició no 100% d'aquests objectius pot produir resultats indesitjats en provocar que la IA actuï de manera perjudicial o inesperada. Per altra banda, a mesura que els sistemes d'AI esdevenen més intel·ligents i generals és cada cop més complicat exercir-ne un control inclús a curt i mig termini, la qual cosa serà encara més difícil si assoleixen l'autonomia suficient per adoptar estratègies no previstes o no aprovades pels humans.

Un aspecte addicional a considerar és la dimensió geopolítica dels riscos associats a la IA. El desenvolupament de sistemes d'IAg avançats s'està convertint en una prioritat estratègica per a moltes potències mundials, amb el potencial de crear noves dinàmiques de poder internacional. La cooperació global és essencial per evitar una carrera armamentística d'IA on els estàndards de seguretat i ètica quedin subordinats a objectius competitius. Per aquest motiu, iniciatives com el Pacte Global sobre la IA, o els treballs del Consell d'Europa i de les Nacions Unides sobre aquesta matèria, són crucials per establir marcs de col·laboració internacional en governança d'IA que garanteixin el desenvolupament segur i beneficiós d'aquesta tecnologia.

[27] https://doi.org/10.1098/rsos.240197

Stuart Russell proposa en el vídeo que va tancar la conferencia IASEAI'25 de Paris tres línies estratègiques per garantir la seguretat de la IA: (i) els models haurien d'incorporar explícitament la incertesa sobre les preferències humanes en comptes de tenir objectius fixos; (ii) limitar les capacitats de la intel·ligència artificial general (AGI) de manera que proporcioni solament informació; i (iii) implementar un registre obligatori i permanent de tots els sistemes avançats d'IA per garantir-ne la traçabilitat i responsabilitat — la IA no hauria de poder-se duplicar, operar anònimament ni esquivar regulacions. Per tant és urgent establir un marc de governança ara, abans que l'AGI no sigui una realitat.

El vídeo de M. Tegmark, *AGI is unnecessary, undesirable & preventable*, subratlla la importància d'establir "línies vermelles" que no s'han de creuar en el desenvolupament de la IA i proposa mecanismes de control més robustos i una supervisió internacional, liderada pels EUA i la Xina, per assegurar que la IA es desenvolupi de manera segura i ètica. El vídeo de Y. Bengio, *Can we get the scientific benefits of AI without the risks of autonomous agents?*, destaca els perills associats amb els sistemes d'IA que operen de manera autònoma, subratllant la necessitat de desenvolupar mecanismes que limitin la seva capacitat d'actuar sense supervisió humana.

14. ¿Quines són les implicacions ètiques de l'ús de LA IAg en la recerca científica?

Resposta: L'ús de la IAg en la recerca científica pot accelerar el procés de revisió de la literatura, generar hipòtesis, i fins i tot proposar, planificar, executar i avaluar tasques i nous experiments, amb una mínima intervenció humana. És per això que tenen implicacions ètiques significatives en plantejar riscos, com ara la generació de cites o dades falses, però creïbles, que podrien comprometre la integritat de la recerca i la consistència de les seves aplicacions pràctiques. A més, l'ús d'aquests models podria accentuar biaixos existents en la literatura científica, si no es gestiona adequadament la informació, perpetuant prejudicis i desigualtats.

També sorgeixen qüestions sobre l'autoria i el reconeixement de la contribució de la IAg en la recerca, ja que la línia que separa el treball humà i el generat per IA es torna cada dia que passa més difusa. És crucial, doncs, establir directrius ètiques clares per a l'ús d'aquests models, incloent-hi la transparència en el seu ús i la verificació rigorosa dels resultats generats per evitar la propagació de dades incorrectes o enganyoses. Això serà encara més necessari quan es posin en acció les capacitats prescriptives de la IA generativa i es posin en marxa els anomenats laboratoris autònoms.

SOBRE ELS BIAIXOS DE LA IA I LA FORMA DE COMBATRE'LS

15. ¿Què són els biaixos en els models d'IA i com s'originen?

Resposta: Els biaixos en els models d'IA es refereixen a tendències o prejudicis sistemàtics en les prediccions o decisions del model, de la mateixa manera que ens referim als biaixos conscients o inconscients dels humans en relació al gènere, classe o raça. S'originen a partir de dades d'entrenament no equilibrades, decisions de disseny del model, i factors humans implicats en la recopilació i etiquetatge de dades. De la mateixa manera que diem que hem de promoure una educació igualitària i inclusiva, també hem d'exigir que els sistemes d'IA siguin entrenats amb valors ètics i de manera inclusiva.

16. ¿Com es poden mitigar els biaixos en els sistemes d'intel·ligència artificial generativa?

Resposta: Per mitigar els biaixos en la IAg és fonamental actuar des del procés inicial d'entrenament fins a la seva implementació. Primer, és essencial seleccionar i curar adequadament les dades, assegurant que siguin diverses, representatives i equilibrades. Això implica realitzar auditories periòdiques per identificar possibles desequilibris o omissions significatives que puguin generar biaixos en els resultats. En segon lloc, cal implementar tècniques específiques durant el desenvolupament i entrenament dels models que regulin la seva equitat, i utilitzar metodologies transparents que facilitin la interpretabilitat dels resultats.

La supervisió i avaluació contínua dels sistemes d'IA, mitjançant sistemes de monitorització en temps real i procediments regulars d'avaluació que combinin eines automàtiques amb revisions humanes, és del tot necessària, ja que permet identificar i corregir immediatament qualsevol desviació detectada. Aquest marc de supervisió i avaluació tot i tenir la dificultat afegida de necessitar recursos computacionals importants, permet detectar i corregir ràpidament els biaixos que puguin sorgir en tot el procés, des de l'entrenament fins a la posada en operació i manteniment del sistema d'IA desenvolupat.

Finalment, cal que les empreses que comercialitzin sistemes d'IA capacitin adequadament els equips responsables perquè comprenguin la naturalesa i l'impacte dels biaixos, fomentant una cultura organitzativa basada en l'ètica i la responsabilitat. També s'han d'implementar marcs reguladors clars que assegurin el registre obligatori i permanent de tots els sistemes d'IA i promoguin una rendició de comptes efectiva mitjançant auditories externes independents. Això implica inversions significatives en recursos computacionals i humans, i un compromís sostingut per assolir sistemes alineats amb els valors humans.

17. ¿Quins són els reptes de la verificació i validació dels resultats generats per la IAg?

Resposta: La verificació i validació dels resultats generats per la IAg en general, i dels textos generats per LLMs en particular, presenta diversos reptes importants. En primer lloc, la naturalesa probabilística d'aquests models fa que, en el cas concret dels LLMs, puguin generar respostes que semblin plausibles però que siguin incorrectes. A més, la complexitat dels models dificulta la comprensió de com s'arriba a una determinada resposta o text, la qual cosa fa que sigui difícil rastrejar i explicar el procés seguit per a decidir la sortida del model o la seva traçabilitat. També hi ha el fenomen de l'anomenada "al·lucinació" o "confabulació", atès que els models poden generar informació que sembli coherent, tot i que no es basi en fets contrastables o verificables.

La verificació independent del material generat i de les fonts de les dades d'entrenament en temps real, és un desafiament significatiu degut a que es generen grans volums de text i caldrien molts recursos computacionals i grans centres de dades per supervisar-los. El fet que els centres de dades més importants estiguin en mans privades que, alhora, són les comercialitzadores dels models d'IA generativa, fa que aquesta verificació independent sigui poc realista.

Cal tenir present que per abordar aquests reptes es necessiten, a més a més, eines avançades de verificació automàtica, sistemes robustos de comprovació de fets, i la integració de coneixements d'experts humans en el procés de validació. També és crucial desenvolupar metodologies transparents que permetin auditar i comprendre el funcionament intern dels LLMs i de la IAg.

18. ¿Quins són els principals reptes en la regulació de la IA generativa?

Resposta: Els principals reptes en la regulació de la IA generativa inclouen:

- La ràpida evolució dels sistemes d'IA en general, i dels LLMs en particular, supera amb escreix la capacitat dels legisladors per regular-los eficaçment i per adaptar les normatives pertinents de manera continuada i efectiva. A més, quan els sistemes esdevinguin més generals i autònoms tindran més a l'abast la capacitat per esquivar el control humà[28].
- La naturalesa global d'Internet, que complica l'aplicació de regulacions nacionals i fa imprescindible una regulació global, en la qual hi participin els governs, els experts, les empreses tecnològiques i la societat en general per tal d'assegurar la seva efectivitat.

[28] Managing extreme AI risks amid rapid progress

- La necessitat de trobar un equilibri entre la promoció de la innovació, la seva comercialització i la protecció dels drets individuals, incloent-hi la privacitat, la seguretat i la llibertat d'expressió.
- La dificultat de definir i mesurar conceptes complexos com la transparència i l'equitat ("*fairness*") en sistemes d'IA generativa que són molt sofisticats.
- La manca d'un marc normatiu global i de la capacitat computacional que permetin una supervisió adequada dels algorismes i dels processos de presa de decisions en temps real o amb un breu temps de resposta.
- La necessitat de formació específica i continuada dels reguladors en matèria d'IA generativa per assegurar que les regulacions es fonamentin en un coneixement profund i actualitzat d'aquesta tecnologia.
- La possibilitat de l'ús malintencionat de la IA, que requereix una regulació que inclogui la previsió i mitigació de tots els possibles abusos.

Cal destacar també el desafiament que representa el que podríem anomenar "bretxa reguladora temporal" o el considerable desfasament entre la velocitat vertiginosa d'adopció d'una tecnologia disruptiva com la IAg i el ritme molt més lent d'implementació de marcs reguladors efectius. Durant aquest període crític, sistemes d'IA potencialment perillosos podrien operar sense la supervisió adequada, creant riscos significatius. Per mitigar aquesta vulnerabilitat, seria necessari desenvolupar mecanismes reguladors capaços d'anticipar-se i d'evolucionar en temps real en resposta a les noves capacitats i riscos emergents de la IA. Cal dir que perquè tota regulació sigui efectiva ha de ser clara, coneguda, verificable, el seu compliment exigit i el seu incompliment sancionat. Ha de ser una regulació que generi confiança. Paral·lelament, resultaria imprescindible invertir en sistemes de registre obligatori i eines de monitoratge continu que permetessin no només detectar sinó també abordar preventivament els riscos potencials abans que poguessin materialitzar-se en conseqüències adverses per a la societat.

SOBRE L'EQUITAT I LA GOVERNANÇA DEMOCRÀTICA

19. ¿Com es pot garantir l'accés equitatiu a la IAg per tal que sigui de tothom i per a tothom?

Resposta: Garantir l'accés equitatiu a la IAg implica superar diverses barreres. En primer lloc, cal reduir la bretxa digital que existeix actualment en molts territoris físics i humans, mitjançant la millora de la infraestructura tecnològica en les àrees més vulnerables o menys desenvolupades tecnològicament. En

segon lloc, és important fomentar el desenvolupament de models en diferents llengües per evitar la marginació de comunitats lingüístiques minoritàries. També cal promoure la sensibilització sobre la IAg perquè la població en general conegui aquesta tecnologia, i dur a terme tasques de formació per augmentar la comprensió i l'ús efectiu d'aquestes tecnologies en els sectors públics i privats. A més a més, caldria consensuar, desenvolupar i implementar polítiques que fomentin la distribució equitativa dels beneficis de la IA, com ara l'accés obert a certs models i aplicacions, no solament a ONGs sinó a ciutadans o comunitats en situació de vulnerabilitat. Finalment, cal considerar les necessitats de les persones amb discapacitats en el disseny i la implementació d'interfícies d'usuari per a aquests sistemes.

20. ¿Com pot la IAg afectar a la democràcia?

Resposta: Els models de llenguatge de gran escala esdevindran un actor més en el processos de diàleg i d'interacció humana, els quals són una part important dels processos democràtics. Per exemple, els LLMs impactaran en la comunicació i el diàleg públic per la seva capacitat de crear continguts amb informació veraç o falsa, i també incrementaran i amplificaran les veus d'aquest diàleg en totes les seves formes i canals, la qual cosa planteja desafiaments en termes de manipulació i seguretat de la informació, sobretot en els processos participatius, com ara els processos electorals. Caldran eines de vigilància i monitoratge efectives, que treballin en línia i en temps real. Per tant, hem de treballar de manera local i global per assegurar la transparència algorítmica i la curació responsable de continguts i de la seva inclusivitat, i alhora facilitar la participació ciutadana en tots els processos democràtics, començant pels que afectin directament a la regulació i legislació de la IA.

21. ¿Com poden influir els sistemes d'IAg la presa de decisions en els sectors públic i privat?

Resposta: Els LLMs poden tenir un impacte profund en la presa de decisions tant en el sector públic com en el privat, atès que poden analitzar ràpidament grans volums de dades, generar resums d'informació i d'informes detallats, i oferir recomanacions basades en patrons identificats a les dades. La IAg pot ajudar al sector públic en l'elaboració de polítiques, en la gestió de la participació ciutadana, en el disseny i execució d'accions com a resposta a consultes ciutadanes, i en la millora de la qualitat i diversitat dels serveis públics mitjançant l'anàlisi de dades socials i econòmiques.

En el sector privat, la IAg pot ser utilitzada per a l'anàlisi de mercats, la presa de decisions estratègiques i la millora de l'eficiència operativa de cada organització. No obstant això, aquesta incorporació de la IA en els processos esmentats,

planteja preocupacions sobre la seva transparència i les responsabilitats que s'hagin d'assumir en cas de conflicte, especialment quan les decisions que se'n derivin tinguin un impacte significatiu en la vida de les persones. També hi ha el risc que els biaixos presents en les dades d'entrenament es reflecteixin en les recomanacions dels models. Per tant, és crucial crear comitès d'ètica i de seguiment que implementin mecanismes de supervisió humana i estableixin els marcs ètics clars per a l'ús de la IAg en la presa de decisions de cada organització, tal com regula la *UE Artificial Intelligence ACT*, publicada el 12 de juliol de 2024[29].

SOBRE L'EDUCACIÓ, L'ART, LA LLENGUA I LA CULTURA

22. ¿Com pot l'ús de la IAg afectar a l'educació?

Resposta: L'impacte de la IAg en l'educació serà significatiu i ràpid, no solament per l'ús extensiu que ja en fan la majoria d'alumnes, des de l'ESO a l'educació superior, sinó també pel fet que els professors hauran de canviar les eines i recursos docents per afavorir els processos d'aprenentatge de caire més constructivistes[30]. Cal tenir present que els LLMs poden oferir assistència personalitzada als estudiants, adaptar-se a les seves necessitats individuals, generar recursos i materials educatius a mida de cada patró d'aprenentatge, i facilitar l'accés a publicacions originals escrites en diferents llengües, ja sigui directament o a través de resums produïts artificialment.

L'ús d'aquests assistents individualitzats d'IAg planteja reptes importants, com ara la possible dependència excessiva (*overreliance*) d'aquestes eines que podria afectar el desenvolupament de certes habilitats essencials dels humans, com el pensament crític, el treball en equip, la capacitat per resoldre problemes i la innovació. Pel que fa als professors[31], l'ús de la IAg pot conduir a la planificació de lliçons que no construeixin efectivament el coneixement dels estudiants, tutories que poden confondre els estudiants amb respostes incorrectes, i materials didàctics basats en conceptes erronis. Davant d'aquest panorama, és essencial que els educadors i les institucions educatives desenvolupin polítiques que assegurin que les eines generades per IA siguin rigorosament avaluades i verificades, i s'integrin de manera ètica i efectiva en el sistema educatiu, per tal de garantir un equilibri entre l'ús de la tecnologia i la necessitat de desenvolupar habilitats humanes en un marc d'estricte respecte als drets fonamentals[32].

[29] https://artificialintelligenceact.eu/the-act/
[30] https://www.wikiwand.com/ca/Constructivisme_(pedagogia)
[31] https://www.cognitiveresonance.net/resources.html
[32] https://rm.coe.int/artificial-intelligence-and-education-a-critical-view-through-the-lens/1680a886bd

No obstant això, ni la manca de polítiques clares ni els reptes plantejats han impedit que, en l'ensenyament superior, s'hagin desenvolupat i avaluat favorablement activitats a l'aula específicament dissenyades per potenciar el pensament crític, principalment en el procés de plantejar preguntes incisives i profundes, d'avaluar informació per extreure conclusions lògiques, i de comprendre temes complexos[33]. Aquestes experiències i d'altres dutes a terme per membres de CIVICAi per potenciar el pensament crític a les universitats, suggereixen que l'ús de la IAg a les aules es podria emmarcar en una metodologia fonamentada en la maièutica[34], amb un format d'ensenyament similar al de l'antiga escola socràtica, sota el lideratge de cada professor.

Un format d'ensenyament obert i participatiu facilitaria la reflexió i el pensament crític, promovent discussions profundes i l'intercanvi d'idees entre estudiants i professors. Amb els estudiants tenint assistents personals intel·ligents a la butxaca, aquest canvi de model podria enriquir l'experiència educativa, fomentar una educació més en col·laboració i centrada en l'estudiant, i promoure sistemes d'avaluació més personalitzats i dinàmics. L'any 2023 es va dur a terme a la universitat de Harvard un estudi pilot, controlat i aleatoritzat, per avaluar l'aprenentatge i les percepcions d'estudiants universitaris quan se'ls presenta el contingut d'una assignatura de física, en l'àmbit de les ciències de la vida, amb un chatbot d'IAg en comparació a l'aprenentatge en classes d'aprenentatge actiu[35]. Els resultats mostren que el tutor basat en la IAg no solament va ajudar els estudiants a aprendre més del doble de continguts en menys temps, sinó que també els va motivar i implicar més en el seu aprenentatge.

Cal aprofitar les tecnologies d'IAg per promoure un context educatiu on la reflexió crítica i el debat intel·lectual siguin centrals, amb la finalitat d'assegurar que els estudiants desenvolupin les habilitats necessàries per verificar, interpretar i utilitzar informació complexa de manera beneficiosa, responsable i ètica. Alhora, s'aconseguiria fer més permeables els verticals de cada assignatura, fer evolucionar l'estructura medieval de les universitats i retornar el coneixement allà on va néixer: al procés de fer preguntes per construir coneixement. Sense oblidar, però, la necessitat que tant els estudiants com els professors entenguin els riscos de la IA, amb la finalitat d'interactuar-hi d'una manera segura, ètica i responsable en l'àmbit educatiu i més enllà[36].

[33] Leveraging chatgpt for enhancing critical thinking skills
[34] https://ca.wikipedia.org/wiki/Mai%C3%A8utica?wprov=sfti1#
[35] https://doi.org/10.21203/rs.3.rs-4243877/v1
[36] UNESCO's AI competency frameworks for students and teachers

23. ¿Poden els LLMs interpretar i comprendre contextos culturals i socials complexos?

Resposta: Els LLMs (Large Language Models) actuals tenen la capacitat d'identificar i generar llenguatge en contextos culturals i socials basats en les dades amb què han estat entrenats. No obstant això, la seva comprensió d'aquests contextos és limitada i superficial, ja que es basa principalment en patrons estadístics i correlacions trobades en grans volums de text.

Els LLMs poden reconèixer i reproduir patrons de llenguatge que són comuns en diferents cultures i situacions socials, però no posseeixen una comprensió profunda o una consciència real dels matisos culturals i socials subjacents. Això significa que, tot i que poden semblar entendre i respondre de manera coherent en moltes situacions, la seva capacitat per interpretar contextos complexos és limitada. Aquesta limitació, tot i les millores en la capacitat de raonament dels models més avançats, es fa especialment evident en situacions que requereixen empatia, sensibilitat cultural o una interpretació contextual més rica. Per exemple, un LLM pot no captar les subtileses d'una conversa que impliqui ironia, sarcasme o referències culturals específiques d'una regió o grup social determinat. A més, els LLMs poden cometre errades o malentesos quan s'enfronten a situacions que no estan ben representades en les seves dades d'entrenament.

Les limitacions de la intel·ligència artificial generativa, com les exposades a la pregunta #6, són també pertinents per respondre aquesta pregunta #23. Els LLMs no tenen experiències pròpies ni la capacitat de sentir emocions o comprendre les emocions dels altres, el que limita la seva capacitat per interpretar i respondre adequadament en contextos culturals i socials complexos.

24. ¿Com poden afectar els LLMs la diversitat lingüística i cultural?

Resposta: Els LLMs poden tenir un impacte significatiu en la diversitat lingüística i cultural. D'una banda, poden ser una eina molt útil per a la preservació de llengües minoritàries. Això es pot aconseguir generant contingut en aquestes llengües i facilitant la traducció automàtica, cosa que ajuda a revitalitzar llengües en perill d'extinció i a mantenir vives les tradicions culturals.

D'altra banda, tenen el risc de reforçar el domini de les llengües majoritàries, com l'anglès, ja que la majoria d'aquests models s'entrenen principalment amb dades en aquestes llengües. Això pot reduir la visibilitat i l'ús de les llengües minoritàries. A més, els LLMs poden influir en la manera com s'expressen les idees en diferents cultures, homogeneïtzant expressions culturals diverses i eliminant matisos importants.

Per mitigar aquests riscs, cal que el desenvolupament d'aquests models inclogui dades diverses, tant culturals com lingüístiques, i que hi hagi una col·laboració estreta amb els agents culturals de les llengües afectades per assegurar un tractament harmònic i respectuós amb totes les cultures. D'aquesta manera, aprofitarem els beneficis dels LLMs sense comprometre la riquesa de la diversitat lingüística i cultural.

25. ¿Com pot contribuir la IAg a la preservació i estudi del patrimoni cultural intangible?

Resposta: Els sistemes d'IAg poden ser eines valuoses per a la preservació i estudi del patrimoni cultural intangible. Poden ajudar a processar i analitzar grans volums de dades culturals, incloent-hi històries orals, cançons tradicionals i pràctiques culturals. Poden assistir en la transcripció i traducció de llengües en perill d'extinció, per tal de facilitar la seva preservació i estudi. També poden generar representacions interactives de pràctiques culturals, fent-les més accessibles i comprensibles per a la seva difusió a un públic més ampli, i també per desenvolupar una major consciència i apreciació del patrimoni cultural en general.

No obstant això, és crucial involucrar les comunitats culturals en aquest procés per garantir que les representacions siguin precises i respectuoses amb les tradicions. Això també ajudarà a abordar qüestions de propietat intel·lectual i consentiment en l'ús de dades culturals sensibles. En qualsevol cas, les comunitats beneficiàries han de tenir el control sobre com es recopilen, utilitzen i difonen les seves tradicions culturals per tal d'assegurar que el patrimoni cultural es preservi de manera ètica i respectuosa.

26. ¿Com afecta o pot afectar la IAg la creativitat artística i la producció cultural, i quines implicacions ètiques, legals i socioeconòmiques s'entreveuen a curt i llarg termini?

Resposta: La IA generativa ha posat en alerta la majoria d'àmbits de l'activitat artística i la producció cultural. Aquests sistemes, capaços de generar música, art visual, literatura i contingut audiovisual, qüestionen els límits de la creativitat humana en oferir fonts d'inspiració alternatives i eines per a la creació artística. La seva capacitat per influir la producció cultural de manera transversal també pot contribuir a reduir les barreres tècniques i a diversificar els recursos creatius. El fet que la IAg pot introduir formes d'art interactiu i personalitzat no solament pot canviar l'experiència artística, sinó també modificar la percepció de l'autenticitat i el valor de les obres artístiques.

Tanmateix, aquestes transformacions també comporten desafiaments significatius. Des del punt de vista ètic i legal, es plantegen qüestions complexes sobre l'originalitat, l'autoria i els drets de propietat intel·lectual de les obres generades per IA. En aquest sentit, és significatiu el fet que més de 1.000 músics s'han unit per llançar un àlbum titulat "*Is This What We Want?*"[37] com a protesta contra les modificacions proposades per la llei de drets d'autor del govern del Regne Unit. L'àlbum inclou gravacions d'estudis i espais de representació buits, simbolitzant l'impacte potencial en els mitjans de vida dels artistes si s'implementen les modificacions proposades.

La IAg pot provocar una reestructuració profunda del mercat laboral en el sector artístic a causa del desplaçament potencial de certs rols creatius i a l'emergència de noves professions. En aquest context, serà crucial fomentar la col·laboració entre artistes humans i la IAg, assegurant que aquesta sigui un complement informatiu i no una substitució de la creativitat humana. També existeix el risc d'una homogeneïtzació de la producció artística, així com de canvis en la valoració econòmica de l'art i la creativitat.

Davant aquests reptes, serà essencial no només desenvolupar marcs ètics i legals que regulin aquestes noves dinàmiques, sinó també investigar i avaluar a llarg termini l'impacte que tindrà la IAg en la diversitat cultural i l'expressió artística. Fomentar una col·laboració equilibrada entre humans i la IAg, i educar el públic sobre les capacitats i limitacions de l'art generat per intel·ligència artificial, serà fonamental per assegurar un futur en que la tecnologia enriqueixi, en lloc de limitar, l'expressió cultural.

Finalment, caldrà garantir la protecció dels drets dels artistes, estudiant possibles conseqüències, implicacions o fins i tot compensacions en aquest nou entorn creatiu. Aquesta revolució ens obliga a plantejar-nos preguntes fonamentals sobre la naturalesa de la creativitat, la preservació del patrimoni cultural, l'evolució de les identitats culturals, i quin futur volem per l'expressió cultural humana en l'era de la intel·ligència artificial generativa, que tot just comença.

SOBRE LA SOSTENIBILITAT I LA SALUT

27. ¿Quines implicacions tenen els sistemes d'IA en el canvi climàtic?

Resposta: Els sistemes d'IA generativa requereixen una potència computacional creixent, la qual cosa causa un impacte ambiental significatiu en exercir una pressió significativa sobre les xarxes energètiques globals, els recursos hídrics i les reserves minerals. Per mitigar aquest impacte cal situar la sostenibilitat al

[37] https://www.isthiswhatwewant.com

centre de les discussions sobre ètica i seguretat de la IA i fomentar una col·laboració global per desenvolupar pràctiques i polítiques que minimitzin els costos ambientals associats amb el desenvolupament i la implementació de la IA. Això implica anar més enllà de les estratègies per reduir el consum energètic associat a aquests models i posar l'èmfasi en el cicle de vida o cadena de subministrament de la IA, des de l'extracció de minerals fins a la gestió dels residus electrònics, sense oblidar l'ús intensiu d'aigua de refrigeració en els centres de dades.

Si analitzem l'impacte energètic dels sistemes d'IAg, ens adonarem que bàsicament hi ha dues estratègies possibles: reduir el consum o diversificar les fonts d'energia utilitzades. En relació al consum, es treballa en el desenvolupament de models més eficients en termes de càlcul (entrenament i operació), en l'optimització dels algorismes per minimitzar els recursos computacionals necessaris, i en l'ús de maquinari especialitzat com xips adaptats als models d'IA generativa. També s'exploren actualment tecnologies emergents, com ara els sistemes computacionals híbrids o analògics, que podrien oferir solucions més eficients en termes energètics. Pel que fa a les fonts d'energia per alimentar els centres de dades, les empreses d'IA incorporen les energies renovables i, sobretot, han adoptat l'estratègia d'adquirir o fusionar-se amb empreses de producció d'energia nuclear. Cal tenir present que l'energia que consumeix la IA generativa actual és superior al consum energètic d'alguns dels 193 estats de l'ONU.

Tanmateix, la IAg, més enllà del seu impacte energètic, pot jugar un paper crucial en la sostenibilitat ambiental i en la lluita contra el canvi climàtic. Aquests sistemes poden optimitzar la gestió de recursos naturals mitjançant el processament d'enormes conjunts de dades ambientals per detectar patrons de degradació, predir escassetat de recursos o identificar zones prioritàries per a conservació. En la planificació urbana sostenible, poden ajudar a dissenyar ciutats més eficients energèticament, planificar rutes de transport optimitzades i simular l'impacte de diferents polítiques urbanes abans d'implementar-les. També poden identificar patrons i tendències, fer prediccions sobre fenòmens meteorològics extrems, com la de la trajectòria d'huracans de manera ràpida i efectiva[38,39], i millorar la precisió dels models climàtics existents. Això podria ajudar a comprendre millor els efectes de les emissions de gasos d'efecte hivernacle i d'altres factors antropogènics.

[38] Just how much can we trust Ai to predict extreme weather
[39] https://www.freethink.com/robots-ai/ai-based-weather-forecasting

En l'àmbit de les energies renovables, la IA pot optimitzar la ubicació, el disseny i l'operació de parcs solars i eòlics, millorar l'eficiència de les microxarxes elèctriques i predir amb precisió la producció energètica per facilitar la integració de renovables a la xarxa. Per la biodiversitat, aquests sistemes poden ajudar a identificar i classificar espècies, monitorar ecosistemes mitjançant l'anàlisi d'imatges de satèl·lit i dades de sensors, i modelar l'impacte potencial de diferents escenaris climàtics en hàbitats específics.

Paral·lelament, aquests sistemes són eines valuoses per a l'economia circular, optimitzant processos de reciclatge, dissenyant productes més sostenibles i identificant oportunitats per reduir residus en processos industrials. A més, la IAg pot ser utilitzada per avaluar l'impacte de diferents polítiques ambientals i oferir recomanacions basades en dades per a una gestió més efectiva del canvi climàtic. Finalment, poden millorar la comunicació i l'educació ambiental, traduint conceptes científics complexos a formats accessibles pel públic general i personalitzant missatges per promoure comportaments i estils de vida més sostenibles.

No obstant això, per maximitzar aquests beneficis, és crucial assegurar que el desenvolupament i desplegament d'aquests sistemes segueixi principis de sostenibilitat, minimitzant la seva petjada ambiental i assegurant que les solucions proposades siguin inclusives i considerin els contextos socials i econòmics on s'apliquin.

El vídeo de K. Crawford, *Hyperscaled: Bridging AI safety, ethics and sustainability* posa en perspectiva el tema de la sostenibilitat i la cadena de subministrament dels sistemes d'IAg, en el context d'ètica i seguretat.

28. ¿Com pot influir la IAg en la detecció i prevenció de crisis de salut pública?

Resposta: La IAg pot ser una eina potent en la detecció i prevenció de crisis de salut pública. La seva capacitat per analitzar grans volums de dades de salut, literatura científica i informes de mitjans i xarxes socials permet identificar patrons emergents que podrien ser indicatius de brots de malalties abans que es converteixin en crisis a gran escala. Aquests models poden contribuir a una resposta més ràpida en situacions d'emergència i millorar la comunicació amb les poblacions afectades, mitjançant la difusió precisa sobre salut pública en múltiples llengües.

Malgrat els avantatges potencials, també s'han de tenir en compte els riscos per un ús inadequat d'aquests models en relació a la privacitat de dades de salut i a la possibilitat de generar falses alarmes. Per aquest motiu, és imprescindible assegurar que les dades utilitzades siguin de qualitat i representin

adequadament la diversitat de la població, i que els models d'IAg s'integrin de manera rigorosa en els sistemes de salut pública, principalment en els serveis d'epidemiologia, amb protocols clars per a la verificació i difusió de tota la informació generada per IA.

29. ¿Com pot millorar la IAg els sistemes de salut, tant des del punt de vista de l'experiència del pacient com de la detecció i tractament de les malalties que puguin patir?

Resposta: La IA generativa pot transformar els sistemes de salut en profunditat, millorant tant l'experiència del pacient com la detecció i tractament de malalties en l'atenció primària, l'especialitzada i en l'hospitalària. Pel que fa a l'experiència del pacient, un aspecte clau és la qualitat de la interacció i la compassió que mostren els professionals de la salut durant les visites presencials. La IAg pot contribuir a alleugerir la càrrega dels professionals en tasques rutinàries, permetent-los centrar-se més en el tracte humà. Per exemple, la IAg pot ajudar en el registre automàtic de la informació del pacient, transcrivint a partir de la seva pròpia veu els motius de la consulta o els símptomes que descrigui. Aquest registre es pot integrar directament a la seva història clínica, sempre després d'una revisió per part del facultatiu, i la IA generativa pot suggerir accions adequades, com ara una derivació a un especialista, un ingrés hospitalari o un tractament a seguir. Aquesta automatització no només milloraria l'eficiència, sinó que permetria als professionals de la salut dedicar més temps a l'atenció directa i empàtica dels pacients, elevant així la qualitat global de l'atenció mèdica.

En termes de detecció i tractament de malalties, els models IAg avançats poden analitzar grans volums de dades multimodals, com imatges mèdiques, registres electrònics de salut i dades de sensors, per identificar patrons que podrien passar desapercebuts per als humans. Això seria especialment valuós en entorns crítics com les Unitats de Cures Intensives (UCI), on l'anàlisi en temps real de dades de fonts diverses pot generar pre-alertes i alertes abans que es produeixi un deteriorament significatiu en la salut del pacient, facilitant una intervenció precoç. Aquestes capacitats poden millorar significativament la gestió del risc i reduir els esdeveniments adversos evitables.

A més, la IAg poden tenir un impacte en la millora de la gestió dels fluxos de treball i dels recursos humans, econòmics i d'equipaments en termes generals, i en els serveis d'infermeria en particular, per la seva capacitat d'analitzar dades històriques i en temps real del sistema de salut. Per exemple, una optimització de la planificació de les jornades laborals que analitzés les càrregues de treball i tingués en compte les habilitats, les preferències i els perfils individuals dels

professionals d'infermeria, permetria reduir els errors humans i identificar oportunitats de millora. Automatitzar tasques repetitives i administratives, com l'entrada de dades, la programació de cites, i el seguiment de medicació, així com millorar la coordinació dels equips, facilitaria una gestió més eficient i augmentaria la seguretat i la qualitat de l'atenció als pacients.

Finalment, l'adopció de la IAg en els sistemes de salut hauria de produir-se mitjançant col·laboració entre sistemes de salut de diferents països. Això permetria compartir de manera segura dades anonimitzades, diagnòstics, tractaments i resultats clínics, la qual cosa acceleraria els avenços mèdics globalment i milloraria l'abordatge de crisis sanitàries a escala mundial. En resum, la integració de la IAg en els sistemes de salut té el potencial de millorar significativament tant l'atenció al pacient com l'eficiència clínica. No obstant això, és crucial garantir un ús ètic i segur d'aquestes tecnologies, protegint la privacitat de les dades mèdiques i assegurant que les decisions automatitzades han estat proposades per agents d'IA que han passat pel cribratge de proves controlades i aleatòries[40], amb la participació i supervisió de professionals mèdics.

SOBRE EL TREBALL: REPTES I DESAFIAMENTS

30. ¿Com pot la IA generativa transformar els llocs de treball?

Resposta: La IAg ha començat a transformar el mercat laboral, afectant tant la naturalesa dels llocs de treball com la distribució dels beneficis econòmics. La seva capacitat per automatitzar tasques cognitives o que requereixin processament de llenguatge, com la redacció de textos, l'anàlisi de documents i la generació de contingut creatiu, afectarà a mig termini tots els sectors productius i professionals com el periodisme, el disseny, l'enginyeria, les finances i els serveis professionals. Això pot augmentar la productivitat, i alhora generar preocupacions sobre l'estancament salarial i la concentració del poder econòmic en les poques empreses tecnològiques que controlen les infraestructures i els models d'IA.

Un dels principals efectes de la IAg pot ser la substitució de treballadors menys qualificats o en funcions que abans requerien habilitats humanes específiques, des de la redacció de textos fins a l'anàlisi financera, pel fet de reduir costos a les empreses. Això pot també limitar les oportunitats laborals per a professionals de qualificació mitjana i dificultar la seva transició cap a nous sectors, i també les de professionals altament especialitzats per l'automatització de processos qualificats d'alt risc.

[40] https://ca.wikipedia.org/wiki/Prova_controlada_aleatòria

D'altra banda, la IAg també pot generar noves oportunitats, ja sigui o bé en àmbits on la creativitat, l'estratègia i la supervisió humana siguin essencials , o bé en tasques relacionades amb el desenvolupament i supervisió de la IA, la gestió de dades i la ciberseguretat, entre d'altres. La seva implementació pot permetre als treballadors centrar-se en tasques més complexes i d'alt valor afegit, sempre que hi hagi una formació i adaptació adequades. No obstant això, sense polítiques que protegeixin els drets laborals i redistribueixin els beneficis de l'automatització, el risc és que la IA accentuï les desigualtats i afavoreixi la concentració del poder en poques corporacions.

Per garantir que la IAg contribueixi positivament a l'economia i al mercat laboral, és fonamental establir mesures de regulació que evitin la monopolització, promoguin una transició justa per als treballadors afectats i assegurin que els guanys derivats d'aquesta tecnologia beneficiïn el conjunt de la societat. Això inclou polítiques de formació i requalificació professional, regulació sobre l'ús ètic de la IA i mecanismes per garantir una distribució més equitativa de la riquesa generada per l'automatització.

El vídeo de J.E. Stiglitz, Ai and Economic Risk: Assessment and Mitigation tracta clarament la problemàtica d'aquesta pregunta i de les següents, en el marc general de l'economia i també de la informació i desinformació.

31. Quins són els efectes dels LLMs en el periodisme i els mitjans de comunicació?

Resposta: Els LLMs ja han transformat el periodisme i els mitjans de comunicació en diversos aspectes, pel fet que poden automatitzar la generació de notícies i articles, i augmentar la rapidesa i l'eficiència en la producció de continguts. També poden ajudar en la investigació periodística mitjançant l'anàlisi de grans volums de dades per identificar tendències i patrons, així com en la verificació de fets abans que siguin noticia. L'ús d'aquests models també planteja riscos, com ara la difusió d'informació no (o poc) supervisada i la transformació massa ràpida del sector periodístic i del perfil professional del periodista. La generació automàtica de continguts no hauria de minvar sinó potenciar el paper dels periodistes per tal de garantir la qualitat i la profunditat dels mitjans de comunicació.

És essencial que aquests mitjans deixin constància de la forma i de l'ús dels LLMs en cada notícia o article d'opinió. També cal que implementin mecanismes robusts de verificació de fets, que mantinguin un equilibri entre l'ús d'IA i la supervisió humana, i que desenvolupin polítiques clares sobre la transparència i l'ètica en l'ús de LLMs. A més, cal fomentar la col·laboració entre experts en IA i

els periodistes per assegurar que els continguts generats siguin precisos, imparcials i de qualitat.

32. ¿Quines són les implicacions de l'ús de la IAg en la creació i gestió de contingut en plataformes de xarxes socials?

Resposta: Les implicacions de l'ús de la IAg en les xarxes socials són moltes, diverses i complexes. La IA generativa pot millorar la moderació de contingut, detectant i filtrant llenguatge ofensiu, discurs d'odi i desinformació de manera eficient. També poden personalitzar el contingut i l'experiència dels usuaris, la qual cosa pot limitar l'exposició a perspectives diverses i reforçar biaixos existents. A més a més, hi ha la possibilitat que es manipuli l'opinió pública amb informació falsa o enganyosa generada a gran escala per la IA.

Per tant, calen polítiques de regulació transparents sobre l'ús de contingut generat per IA, desenvolupar mecanismes robustos de detecció de *deepfakes* i desinformació, i educar els usuaris sobre la presència i les limitacions del contingut generat per la IA en aquestes plataformes. També és important fomentar la col·laboració entre les plataformes de xarxes socials, els reguladors i la societat civil per abordar aquests reptes de manera efectiva.

Un aspecte addicional a considerar és l'impacte de la IAg en la formació i percepció de la identitat digital. Amb la proliferació de continguts generats per la IAg, la distinció entre expressions humanes genuïnes i continguts artificials es difumina, cosa que pot alterar profundament com construïm i interpretem les identitats en entorns digitals. Això planteja qüestions fonamentals sobre l'autenticitat, la confiança i la veracitat en les comunicacions en que intervingui la tecnologia, i sobre com poden evolucionar les normes socials i les relacions interpersonals en un context on la IAg participi activament en la producció cultural i el diàleg social.

33 ¿Quins són els desafiaments tècnics més grans en el desenvolupament dels sistemes d'IAg?

Resposta: Els desafiaments tècnics que hauran de superar els desenvolupadors dels models d'IA generativa per fer-los evolucionar cap a una intel·ligència de caire més general, es poden identificar i classificar en funció de la possibilitat que es produeixin, si és que ho fan, a curt (1-2 anys), mig (més de 2 anys) i llarg termini (més de 4 anys). Parlem de possibilitats i terminis d'una manera molt laxa doncs en sistemes complexos, no-lineals, i en ràpida evolució la predictibilitat és baixa.

- Curt termini (1-2 anys): Millora dels algorismes i les arquitectures de hardware per reduir el temps i els recursos necessaris per entrenar i

executar els LLMs; reducció del consum energètic i de la corresponent petjada de carboni (millora del cicle de vida o cadena de subministrament de la IA); millora de la interpretabilitat dels LLMs; millora de la gestió de les dades d'entrenament; i adaptabilitat a dominis o àmbits específics de coneixement, sense perdre informació general.

- Mig termini (més de 2 anys): Multimodalitat avançada per integrar eficaçment en un sol model d'IA generativa diferents modalitats d'entrada i sortida (text, imatge, àudio i vídeo); aprenentatge continu sense necessitat de re-entrenament complet; millora de la capacitat per realitzar raonaments complexos i abstractes, més enllà de la simple associació estadística; incorporació de sistemes de seguretat avançats per protegir la privacitat de les dades i prevenir l'ús malintencionat; i personalització sense comprometre l'eficiència.
- Llarg termini (més de 4 anys): Comprensió contextual integral que doti a la IA generativa de comprensió profunda i dinàmica del context cultural, temporal i específic de la situació; aprenentatge autònom i en temps real, sense intervenció humana; raonament causal per entendre i modelar relacions complexes; integració de la IA connexionista amb la IA simbòlica o d'altres sistemes cognitius per crear sistemes d'IA híbrids que puguin emular aspectes més amplis de la cognició humana; desenvolupament de nous models acoblats amb la computació quàntica o neuromòrfica per millorar l'eficiència computacional i energètica; incorporació als models generatius de mètodes que assegurin un alineament amb els valors humans; i el desenvolupament de l'AGI (Intel·ligència Artificial General) amb tot el que pot comportar quant a integració de capacitats, flexibilitat cognitiva, comprensió contextual profunda, metacognició, nivells d'autoconsciència, entre d'altres desenvolupaments avançats.

ANNEX B. GLOSSARI BÀSIC

TERMINOLOGIA RELACIONADA AMB LA IA GENERATIVA O AMB ALGUNES DE LES SEVES FUNCIONS O CAPACITATS

Consideracions prèvies. Quan parlem de les capacitats i prestacions de la IA generativa ens referim a un conjunt de capacitats descriptives, predictives i prescriptives que permeten dur a terme tasques, com ara classificar, veure-hi, predir tendències, reconèixer patrons, extracció d'informació, aprendre, prendre decisions per a assolir objectius, analitzar xarxes socials, etc., que de manera holística i integrada porta a terme un sol sistema computacional. A més de descriure i predir, cada cop pren més rellevància el desenvolupament de sistemes que tinguin la capacitat prescriptiva i poder prendre decisions de manera autònoma, atès que això facilitaria la posada en marxa d'unitats, departament o laboratoris autònoms que poguessin planificar, executar i avaluar tasques o experiments amb una mínima intervenció humana. La prescripció esdevindrà, per tant, una característica cabdal en l'evolució dels sistemes d'IA actuals.

Abans d'aparèixer el *ChatGPT* 3.5 el 30 de novembre de 2022, les capacitats de classificar i predir s'aconseguien de manera separada per algorismes singulars dissenyats per efectuar de la manera més eficient possible cadascuna d'aquestes accions amb instruccions ben definides. Per tant, tot i que cap d'aquests algorismes singulars pot ser considerat "intel·ligent" en el context i conjunt d'aquest glossari, se'ls ha inclòs perquè alguns dels seus principis o fonaments i objectius de les seves instruccions formen part dels sistemes d'IA generativa actuals.

Glossari

Adulació servil en IA (*sycophancy* en IA): És el comportament que podria tenir la IAg per sintonitzar-se amb els estats emocionals dels humans, d'una manera que, en qualsevol procés d'interacció, no solament reconegués les seves emocions i inseguretats sinó que també hi empatitzés de maneres complexes i subtils, amb la finalitat d'aconseguir la seva confiança o fins i tot una dependència que obrís la porta a possibles manipulacions.

Agents intel·ligents: Entitats autònomes que poden percebre el seu entorn, raonar, aprendre i prendre decisions (actuar) per assolir objectius específics a

partir de la informació rebuda.
https://www.wikiwand.com/ca/Agent_intel%C2%B7ligent

Algoritmes: Conjunt d'instruccions inequívoques que un sistema, i en particular la IA, executa per dur a terme tasques específiques, mesurables i repetibles d'acord amb regles definides..
https://www.wikiwand.com/ca/Algorisme
¿Què és un algoritme?

Algoritme de caixa negra: Algoritme el funcionament intern del qual és difícil o impossible d'entendre, d'explicar o d'examinar. Aquests algoritmes són sovint complexos i poden prendre decisions o fer prediccions sense que es puguin explicar clarament com s'ha arribat a aquests resultats, atès que funcionen com una caixa negra.

Algoritmes d'optimització: Conjunt d'algoritmes per resoldre problemes de minimització o maximització d'una funció objectiu. En situacions de la vida quotidiana això pot consistir en minimitzar o reduir a la mínima expressió pèrdues econòmiques o en maximitzar guanys econòmics en un procés o activitat domèstica o industrial. En llenguatge més abstracte minimitzar significa assolir el valor més petit possible de l'error o desviació de la solució obtinguda (prediccions de l'algoritme) respecte a un conjunt de dades determinades. L'objectiu i funcionalitat d'aquests algoritmes és trobar la millor solució, definida prèviament amb un conjunt de criteris, d'entre totes les solucions possibles.

Algoritmes evolutius: Família d'algoritmes d'optimització inspirats en la teoria de l'evolució, que utilitzen mecanismes com la reproducció o l'herència, la selecció, l'encreuament o recombinació i la mutació per trobar solucions òptimes. Els algoritmes genètics son els més coneguts dels algoritmes evolutius doncs s'inspiren en els mecanismes de l'evolució biològica.

Alineament d'IA: Àrea de recerca dedicada a garantir que els sistemes d'IA actuïn d'acord amb les intencions i els valors humans, fins i tot quan els permetem optimitzar objectius o operar amb un cert grau d'autonomia. Les tècniques actualment més consolidades inclouen l'aprenentatge per reforç amb retroalimentació humana (RLHF), la IA constitucional i les variants d'alineament deliberatiu (vegeu entrades pròpies). Paral·lelament, la disciplina de la interpretabilitat mecanística —que mira d'identificar quins circuits interns d'un model expliquen quins comportaments— ha esdevingut un dels debats centrals

del camp: sense entendre per què un model decideix el que decideix, qualsevol garantia d'alineament és, en darrera instància, conductual i no estructural.

Al·lucinació (o confabulació): Fenomen en què els models d'IAg produeixen contingut que sembla plausible i coherent però que és objectivament incorrecte, inventat o sense base en dades reals. Aquest comportament es produeix especialment quan els models tracten temes poc representats en les seves dades d'entrenament o quan s'enfronten a preguntes ambigües.

Anàlisi de sentiments: Tècnica de processament del llenguatge natural (PLN) que s'utilitza per determinar l'opinió, sentiment o actitud expressada en textos, o a partir de patrons de comportament. S'utilitza àmpliament en l'anàlisi de les xarxes socials o en l'estudi de la satisfacció de clients. https://www.wikiwand.com/ca/An%C3%A0lisi_de_sentiment

Anàlisi de xarxes socials: Estudi de les relacions i interaccions entre actors (persones, organitzacions, etc.) en xarxes socials, mitjançant l'escalat multidimensional i el "*block-modelling*" per identificar grups sobre la base de l'equivalència de les estructures de relacions. Aquestes propostes varen ser implementades mitjançant tècniques de teoria de grafs i estudiar empíricament les xarxes socials.

Aprenentatge actiu: Estratègia d'aprenentatge automàtic on el model d'aprenentatge selecciona/tria activament les dades d'entrenament, de les quals aprèn, de manera que continguin la més i millor informació per millorar el seu rendiment o capacitat de predicció o de reconeixement de patrons. D'aquesta manera el model d'aprenentatge obtén un rendiment més alt en triar les dades pel seu aprenentatge. El procés s'inicia amb un subconjunt petit d'exemples d'entrenament ben definits, el qual s'amplia progressivament i cíclica, amb els exemples que el model és incapaç de predir correctament. D'aquesta manera el model utilitza pel seu aprenentatge solament el subconjunt de dades que li cal per predir o "explicar" tot el conjunt de dades.

Aprenentatge automàtic (*Machine Learning* en anglès - ML): Procés mitjançant el qual un sistema computacional pot aprendre i millorar el seu rendiment a mesura que se li proporciona més dades d'entrenament. Aquest procés utilitza algoritmes o models estadístics per dur a termes tasques determinades d'anàlisi de dades, d'extracció d'informació o d'identificació de patrons, sense que

necessàriament hagin estat explícitament programats per fer-ho. Els algoritmes d'aprenentatge automàtic es poden classificar en les següents categories:

- Aprenentatge supervisat: Els models s'entrenen amb dades etiquetades per predir sortides a partir d'entrades noves. Per exemple, un algoritme d'aprenentatge supervisat pot ser entrenat per reconèixer objectes o subjectes determinats en fotografies o vídeos.
- Aprenentatge no supervisat: Utilitza dades sense etiquetar per trobar patrons, agrupacions o relacions en les dades. Un exemple seria un algoritme que agrupés textos segons la temàtica tractada.
- Aprenentatge semi-supervisat: Combina l'ús de dades etiquetades i no etiquetades per millorar el rendiment del model.
- Aprenentatge per reforç: Els models aprenen a través de la interacció amb el seu entorn i reben recompenses o penalitzacions segons les seves accions. És un aprenentatge a partir de l'experiència que maximitza la recompensa acumulada. S'aplica en l'aprenentatge de jocs.
- Aprenentatge federat: Diversos dispositius o servidors col·laboren per entrenar un model comú sense compartir les seves dades originals, protegint així la privadesa dels usuaris. Un servidor central agrega els models entrenats localment per cada dispositiu amb les seves dades locals, i reenvia aquest model global a cada dispositiu per a ser refinat amb més dades locals. Aquest procés es repeteix fins que el model global deixa de millorar significativament.
- Meta-aprenentatge: Consisteix en aprendre a aprendre per tal de millorar la capacitat d'un sistema per aprendre noves tasques de manera més ràpida i eficient. S'aplica en l'aprenentatge a partir de molts pocs exemples (*few-shot learning*), on un model aprèn a realitzar una nova tasca amb molt poques mostres o dades d'entrenament. El cas extrem d'aprenentatge a partir d'un sol exemple s'anomena *one-shot learning*.

Aprenentatge per reforç amb retroalimentació humana (RLHF): Tècnica que combina l'aprenentatge per reforç amb la valoració humana per millorar els models d'IA. Els humans proporcionen retroalimentació sobre les respostes del model, valorant quines són preferibles, i aquesta informació s'utilitza per ajustar el comportament del model. Aquesta tècnica ha estat crucial per alinear els LLMs amb les preferències i valors humans, i per reduir les respostes nocives o inadequades.

Aprenentatge per transferència: Tècnica que permet utilitzar un model entrenat en una tasca com a punt de partida per entrenar un altre model en una tasca similar o relacionada.

Aprenentatge profund (*Deep Learning*): Subcamp de l'aprenentatge automàtic que utilitza xarxes neuronals amb múltiples capes (xarxes neuronals profundes) per aprendre representacions jeràrquiques del conjunt de dades. S'utilitzen en el reconeixement de veu, la conducció autònoma, etc., i ha revolucionat el processament del llenguatge natural. El models més comuns d'aprenentatge profund son:

- Xarxes Neuronals Recurrents (RNN): Ideals per a dades seqüencials com el text, on l'ordre de les paraules és important. Les RNN tenen la capacitat d'utilitzar la informació d'entrades anteriors per processar les entrades actuals.
- *Long Short-Term Memory* (LSTM): Tipus especial d'RNN que pot aprendre dependències a llarg termini.
- *Transformers*: Model que utilitza mecanismes d'atenció per assignar un pes que determini la importància de diferents paraules en la comprensió del context d'una frase. Aquest model de xarxa neuronal permet el paral·lelisme en l'atenció, la qual cosa ha fonamentat l'èxit en tasques de processament de llenguatge natural.
- BERT (*Bidirectional Encoder Representations from Transformers*): Model pre-entrenat que pot ser afinat per a una àmplia gamma de tasques de processament de llenguatge natural, incloent-hi el reconeixement d'entitats nomenades, la resposta a preguntes, i la classificació de text. Aquest model és únic pel fet de ser entrenat bidireccionalment, el que significa que es té en compte el context de les paraules tant a l'esquerra com a la dreta d'una paraula donada.

Arbres de decisió: Model d'aprenentatge supervisat que representa decisions en forma d'arbre, amb nodes de decisió i fulles que representen les sortides del model.

Atenció (en xarxes neuronals): Mecanisme que permet a una xarxa neuronal focalitzar-se en parts específiques de la informació o dades d'entrada mentre processa seqüències més grans d'aquesta informació.

Autoencoders: Tipus de xarxa neuronal formada per un codificador i un descodificador, que s'utilitza normalment per aprendre representacions

compactes i eficients de les dades d'entrada. Son utilitzats per reduir la dimensió de les dades mantenint les característiques més rellevants (mínim número de variables per explicar el màxim d'informació continguda en un conjunt de dades), eliminació de soroll, i detecció de frau o funcionament deficient d'un equip o sensor. Els *autoencoders* variacionals (VAEs) son un tipus d'autoencoder que formen part de l'aprenentatge automàtic no supervisat, y que son especialment utilitzats en la generació de dades noves i similars a un conjunt de dades existent, com imatges o textos. Els VAEs són diferents dels autoencoders tradicionals perquè, en lloc de comprimir i descomprimir les dades exactament, els VAEs aprenen a representar les dades d'una manera probabilística, el que els permet generar noves dades de manera més natural i diversa.

Biaix en IA: Es refereix a les desviacions sistemàtiques i repetitives en els resultats d'un sistema d'IA que condueixen a una injustícia sistemàtica o discriminació d'alguns individus o grup d'individus degut a decisions inapropiades del sistema. Aquests biaixos es produeixen sovint en sistemes que impliquen l'aprenentatge automàtic, ja que aquests sistemes aprenen a prendre decisions basant-se en les dades amb les quals s'entrenen. Si aquestes dades estan esbiaixades d'alguna manera, és probable que el sistema aprengui aquests biaixos i els perpetuï. També poden ser causats per un disseny inadequat de l'algorisme. Cal ser transparents sobre les limitacions dels algoritmes, i supervisar-los i actualitzar-los contínuament per mitigar qualsevol biaix.

Hi ha diferents tipus de biaixos que poden afectar els algoritmes, segons el seu origen:

- Biaix de dades: Es produeix quan les dades utilitzades per entrenar un algoritme estan esbiaixades en no representar amb precisió la diversitat del sistema que es vol modelar, descriure o predir.
- Biaix de selecció: Es produeix quan la mostra utilitzada per entrenar l'algoritme no és representativa del sistema que es vol modelar, descriure o predir.
- Biaix de Confirmació: Aquest es produeix quan un algoritme està dissenyat d'una manera que recolza biaixos o creences preexistents.
- Biaix en el Disseny de l'Algoritme: El disseny mateix de l'algoritme pot introduir un biaix, com ara la elecció de les característiques utilitzades en un model predictiu o la manera en què l'algoritme tracta certs tipus de dades.
- Biaix en la Interpretació: Fins i tot si l'algoritme i les seves dades no estan esbiaixats, es pot produir un biaix segons com s'interpretin els seus resultats.

Bosc aleatori (*Random Forest*): Mètode d'aprenentatge automàtic supervisat que combina múltiples arbres de decisió, cadascun d'ells entrenat amb una mostra aleatòria de les dades d'entrenament mitjançant un subconjunt aleatori de característiques de les dades en cada node de decisió, per obtenir millor rendiment i evitar que es produeixi un sobre-entrenament de l'algorisme. S'utilitza tant per a tasques de classificació con de regressió.

Bretxa reguladora temporal: Interval de temps entre l'aparició d'una tecnologia innovadora, com la IAg, i la implementació de regulacions adequades per governar-la. Durant aquest període, poden sorgir riscos significatius a causa de la manca de supervisió i de marcs reguladors efectius. Els models de *machine learning* tradicionals tenen una arquitectura fixa després de l'entrenament. Un cop entrenat, el model realitza un nombre determinat d'operacions per a cada entrada, independentment de la complexitat de la tasca.

Càlcul en temps d'inferència o d'operació (*test-time compute*): És la quantitat de recursos computacionals (temps, energia i recursos computacionals) que necessita un model d'intel·ligència artificial quan s'està fent servir, és a dir, quan rep una entrada i ha de donar una resposta — com generar text, resoldre un problema, classificar una imatge, etc. El *test-time compute* permet als models més avançats adaptar el seu procés de resposta o d'inferència a la complexitat del problema, de manera similar a com un humà dedicaria més temps i esforç a un problema complex que a un de simple.

Calibratge d'un model. Procés d'ajustar un algorisme per tal que les seves prediccions coincideixin, en termes de probabilitat, amb les freqüències observades o reals. Això és crucial en aplicacions d'IA on la confiança en les prediccions és important, com en diagnòstics mèdics o decisions financeres.

Capsule Networks: Son un tipus d'arquitectura de xarxa neuronal proposada per Geoffrey Hinton i col·laboradors que organitza les neurones en grups anomenats càpsules, les quals treballen conjuntament per detectar patrons específics i les seves propietats (com la posició, l'orientació, i l'escala) dins de les dades d'entrada. Aquestes xarxes permeten superar les limitacions que tenen les xarxes neuronals convolucionals (CNN) per gestionar eficaçment les posicions i orientacions dels objectes dins d'imatges, motiu pel qual són especialment útils en tasques de reconeixement d'imatges.

Chatbots: Programes informàtics basats en IA generativa que han estat dissenyats per interactuar o comunicar-se amb els éssers humans a través del llenguatge natural, ja sigui de text o de veu, i realitzar tasques específiques, com ara respondre preguntes o planificar un viatge de plaer o negocis. Utilitzen tècniques avançades de processament de llenguatge natural (NPL) i d'aprenentatge automàtic per respondre a les consultes de manera coherent i contextual. Els *chatbots* més avançats poden mantenir una comunicació bidireccional personalitzada segons l'historial de les interaccions i preferències de l'usuari, son multimodals i multifuncionals, tenen escalabilitat per gestionar múltiples converses simultàniament i de manera multilingüe, poden integrar-se a diferents sistemes d'informació, BBDD o CRM, aprendre de manera contínua i inclús detectar l'estat emocional de l'usuari.

Cibernètica: És una disciplina científica i interdisciplinària que estudia els sistemes de control i la comunicació en màquines i organismes vius, així com les interaccions entre ells. Les pantalles tàctils dels telèfons intel·ligents son un exemple d'element cibernètic d'aquests dispositius. També ho son els sistemes de control en edificis intel·ligents, els d'assistència a la conducció en vehicles moderns o les pròtesis d'extremitats que responen a senyals neuronals.

Ciència de les dades: Disciplina que combina principis i mètodes de diverses àrees com les matemàtiques, l'estadística, la informàtica, i l'expertesa i la comprensió profunda d'un àmbit particular o sector d'activitat per extreure coneixements o informació valuosa de dades d'aquest àmbit o sector. Aquest coneixement és important perquè, un cop processades les dades, permet interpretar correctament les dades, identificar mancances, seleccionar metodologies adequades i validar resultats quan siguin la base per a prendre decisions, identificar patrons i tendències, o desenvolupar productes o serveis.

CIVICAi: Creada el març de 2023 a Catalunya, és la primera associació que defensa els interessos de la ciutadania davant la intel·ligència artificial (IA) i, per tant, té com a objectiu principal aconseguir que la ciutadania participi en la governança de la IA, juntament amb la indústria, l'acadèmia i els reguladors. L'associació està formada per aproximadament 500 membres que treballen, tant a nivell local com global, per aconseguir que la integració de la IA dins la societat sigui harmònica, ètica i pel bé col·lectiu. Té el suport d'un consell social format per més de 30 entitats representatives del món professional, empresarial, i universitari.

Classificació (*clustering*): És una tècnica d'aprenentatge automàtic (ML), supervisat o no supervisat que s'utilitza per agrupar dades (ens, objectes o vectors) en les categories, classes o clústers definides prèviament o automàticament durant el procés de classificació, en funció d'una o més propietats o de relacions intrínseques del conjunt de dades (etiquetes). D'entre les tècniques més habituals de classificació podem esmentar els arbres de decisió, els boscos aleatoris, K-means, SVM, etc.

Classificació de textos: Tasca del processament del llenguatge natural que assigna una o més categories predefinides a un text segons el seu contingut i característiques lingüístiques. Permeten categoritzar textos de manera automàtica per organitzar, filtrar o organitzar grans volums d'informació textual. S'utilitzen tres tipus de classificació. La binària (ex. *spam* o no *spam*), multiclasse que assigna el text a una sola categoria o classe (ex. classificació de notícies per seccions d'un diari digital on cada notícia només pot pertànyer a una secció principal), i multietiqueta que assigna múltiples categories a un sol text (ex. classificació de pel·lícules en plataformes d'*streaming* en diferents gèneres simultàniament). S'utilitzen des de models més tradicionals d'aprenentatge automàtic (ex. SVM) fins als d'aprenentatge profund (ex. *Transformers*).

Comprensió semàntica de la IA generativa: Procés que podria dur a terme un sistema d'IA generativa per comprendre el contingut dels textos que genera, a partir de l'anàlisi del significat de les paraules i la seva relació en el context d'un text. No està demostrada la capacitat dels sistemes d'IA actuals per comprendre els textos que generen tot i presentar algunes incipients propietats emergents.

Comprensió sintàctica de la IA generativa: Anàlisi de l'estructura gramatical de les frases per part dels sistemes de IA generativa. Aquesta capacitat si que la posseeixen els sistemes d'AI generativa actual en generar textos d'una qualitat sintàctica comparable a la d'un humà culte.

Computació afectiva: Camp interdisciplinari que tracta de dotar a les màquines de la capacitat de reconèixer, interpretar i expressar emocions. Combina elements d'intel·ligència artificial, psicologia, neurociència i ciències cognitives. Utilitza tecnologies d'aprenentatge profund, visió per computador, processament de llenguatge natural i sensors biomètrics. Té els desafiaments de captar i aprendre la variabilitat cultural en les expressions emocionals, de respectar la privacitat, de tenir ètica en la detecció de les emocions, de ser

fiable en entorns reals i ser consistent en la gestió de la complexitat i subtileses de les emocions humanes.

Computació de reservori (*Reservoir computing*): Ús d'una xarxa de nodes interconnectats per processar informació de manera dinàmica o en funció del temps. Una part de la xarxa, anomenada "*reservoir*", es manté fixa mentre que només es formen les connexions de sortida per processar informació temporal de manera eficient, la qual cosa és útil per a tasques com el reconeixement de patrons i la predicció de sèries temporals.

Computació en el núvol: Model de prestació de serveis informàtics que permet accedir, sota demanda, a un conjunt compartit de recursos computacionals que es poden configurar (com ara xarxes, servidors, emmagatzematge de dades, aplicacions o programaris i serveis) a través d'Internet. Els models de servei son del tipus "Infraestructura com a servei" (IaaS) que proporcioni recursos de computació, "Plataforma com a Servei (PaaS) que ofereixi un entorn per programar, executar i gestionar aplicacions, i "Software com a Servei" (SaaS) que proporcioni accés a programaris a través d'Internet.

Computació evolutiva: Família d'algoritmes d'optimització inspirats en processos biològics com l'evolució i la selecció natural.

Consciència en la IA generativa: Els sistemes actuals no son capaços d'autocontrolar-se ni de fixar els seus objectius, ni d'integrar inputs sensorials obtinguts continuadament a través de la interacció sensorial amb l'entorn a través d'elements autònoms o sensors, ni de tenir experiències subjectives, ni aprendre a partir dels continguts emergents originals que els mateixos sistemes generin. Per tant, podem afirmar que no tenen consciència. Quan, a més a més de les atribucions anteriors, tinguin memòria i emocions, podrem dir que hauran desenvolupat el que anomenaríem consciència artificial digital, la qual serà col·lectiva i general per naturalesa i, per tant, diferent a la consciència humana.

Contingut generat per IA: Contingut creat o modificat per sistemes d'intel·ligència artificial, com ara imatges, vídeos, textos i música.

Control adaptatiu: Tècniques de control que ajusten dinàmicament els paràmetres d'un sistema per adaptar-se a canvis en l'entorn o als de les condicions de funcionament.

Dades massives (*Big Data*): Conjunt de dades de gran volum, velocitat i varietat

que requereixen tècniques i tecnologies específiques per a la seva anàlisi i processament.

Dades personals: Dades o informació que identifica un individu o persona, la qual n'ha de ser propietària universal. La seva propietat ha d'estar garantida i el seu ús protegit.

Descens del gradient: Mètode d'optimització per ajustar de manera iterativa els paràmetres d'un model connexionista (xarxes neuronals) d'IA fins a obtenir el patrons desitjats de sortida del model en funció de les dades d'entrada. El mètode consisteix en definir primer uns funció que permeti avaluar l'error o diferència entre les dades d'entrada i les prediccions de sortida (funció de pèrdua). Aquesta funció es minimitza iterativament, mitjançant l'actualització dels paràmetres del model, de manera que la funció de pèrdua segueixi la direcció de canvi màxim (la del gradient negatiu) fins que obtenim els resultats desitjats a la sortida de la xarxa neuronal (veure retropropagació o *backpropagation*).

Desinformació sintètica: Contingut fals o enganyós generat mitjançant IA amb la intenció de manipular l'opinió pública o influir en processos democràtics. Inclou *deepfakes*, generació d'articles falsos, i manipulació d'informació que sembla autèntica però que ha estat fabricada artificialment.

Detecció comprimida (*compressed sensing*): Tècnica per a la recuperació o reconstrucció de senyals a partir de només unes poques mesures o dades. Això es fa mitjançant l'explotació del fet que la majoria de dades o bé son zero o bé tenen valors molt petits (esparsitat dels senyals), la qual cosa permetent obtenir imatges o dades amb menys mostres. Això és útil en situacions en que és difícil o costós obtenir mesures completes, com ara en imatges mèdiques o en processos de compressió de dades.

Emergència de capacitats: Fenomen pel qual els models de gran escala (especialment els LLMs) desenvolupen capacitats no previstes i no explícitament programades quan s'escalen en grandària i complexitat. Es caracteritza per l'aparició sobtada de noves habilitats o comportaments que no eren presents en models més petits o més simples.

Enginyeria del coneixement: Disciplina que tracta de la creació, representació, manipulació i adquisició de coneixement en sistemes d'intel·ligència artificial.

Estàndards i normatives en IA: Conjunt de regles, principis i pràctiques establertes per organismes reguladors o professionals per assegurar la qualitat, la seguretat, la privadesa i l'ètica en el desenvolupament i la implementació de la intel·ligència artificial. Podeu trobar una descripció pràctica sobre com ens impactarà el Reglament (UE) 2024/1689 del Parlament Europeu a: https://www.eixdiari.cat/opinio/doc/112416/sobre-el-nou-reglament-de-la-ia.html

Ètica en IA: Estudi i aplicació de principis ètics (morals i socials) en el disseny, implementació i l'ús de sistemes d'intel·ligència artificial, de manera que el seu funcionament sigui responsable, just i beneficiós per la societat. Això implica que tots i cadascun dels processos que sustenten la IA siguin transparents, explicables, auditables, equitatius, respectuosos amb la privacitat i subjectes a responsabilitat civil. Calen regulacions governamentals, directius ètiques d'organitzacions internacionals, codis de conducte corporatiu, i la creació d'una agència global d'IA, accions totes elles que haurien de sustentar-se en un diàleg entre indústria, acadèmia, reguladors i societat en general.

Experiència subjectiva: Conjunt de vivències i percepcions internes que un individu experimenta de manera personal i directa. Aquestes experiències són úniques per a cada persona i inclouen pensaments, emocions, sensacions i impressions que no són directament observables ni poden ser contrastades per altres persones. En el context de la IA, l'experiència subjectiva es refereix a la capacitat que podrien aconseguir les màquines per tenir una consciència interna similar a la dels humans, és a dir, la capacitat de tenir experiències pròpies i autònomes. Els sistemes d'IA generativa actuals son algorísmics, utilitzen correlacions estadístiques i el reconeixement de patrons de grans conjunts de dades d'entrenament que els humans i la Internet els hi han proporcionat, no tenen sensors que els connectin directament i de manera continuada amb l'entorn, les qual coses els incapaciten per tenir experiències subjectives i consciència com les dels humans.

Explicabilitat de la IA (XAI): Capacitat d'un sistema d'IA per explicar els seus processos, decisions i prediccions de manera comprensible per als humans. Les tècniques d'IA explicable permeten entendre com i per què un sistema d'IA ha arribat a una determinada conclusió, facilitant la transparència i la confiança en aquests sistemes. En altres paraules, és la habilitat de fer transparent la caixa negra que representa un model d'aprenentatge automàtic complex.

Extracció d'informació: Processament de dades o de textos per extreure informació útil, com ara patrons, relacions, esdeveniments o fets.

Filtratge en col·laboració: Mètode de recomanació que utilitza les preferències i valoracions d'uns usuaris per tal de predir les preferències d'altres usuaris similars. L'encert d'aquest filtratge depèn de com s'estableixin els criteris de la similitud entre usuaris.

Funció d'activació: Funció utilitzada per una xarxa neuronal per transformar la suma ponderada de les entrades (*inputs*) a cada neurona en una sortida no lineal. En les neurones humanes aquest procés consisteix en el procés biològic de naturalesa electroquímica a través del qual una neurona decideix quina informació o senyal elèctric transmet a les neurones amb les quals està connectada a través de les sinapsis. Les entrades i sortides poden ser inhibidores o excitadores. L'activació d'una neurona humana depèn del seu potencial de repòs, dels senyals d'entrada rebuts a través de les sinapsis amb d'altre neurones, de la combinació d'aquests senyals, de la despolarització de la membrana cel·lular de la neurona afectada, el potencial d'acció o impuls elèctric que es transmet per l'axó de la neurona, de la restauració del potencial de la membrana i de la refractarietat o període d'espera que assegura que els senyals elèctrics viatgin en una sola direcció.

Les neurones o nodes d'una xarxa digital són unitats computacionals més simples, que tenen un nombre molt més limitat de connexions, els pesos de les quals s'ajusten durant l'entrenament, i que segueixen regles matemàtiques molt més simples que les respostes bioelèctriques i bioquímiques de les neurones humanes. En aquest cas, la funció d'activació és una transformació no lineal que una neurona artificial aplica a la suma ponderada de les seves entrades per generar la sortida. És el mecanisme que permet a la xarxa aprendre relacions complexes: sense no-linealitat, qualsevol apilament de capes equivaldria a una sola operació lineal.

Funció de pèrdua: Mesura de l'error entre les prediccions d'un model i les dades reals, amb la finalitat d'optimitzar els paràmetres del model.

Generative Pre-trained Transformer (GPT): Model de llenguatge basat en l'arquitectura *transformer* que pot generar text coherent i realista a partir de dades d'entrenament, mitjançant mecanismes d'atenció que assignen un pes per determini la importància de diferents paraules en la comprensió del context

d'una frase. Per a més informació sobre l'arquitectura subjacent, veure també l'entrada "Transformers".

Governança de la IA: Conjunt de pràctiques, polítiques, normes, i legislacions que regulen el desenvolupament, la implementació i l'ús de la intel·ligència artificial, amb l'objectiu de garantir-que el seu desenvolupament i ús siguin ètics, segurs, i transparents, i contribueixin al bé col·lectiu.

GPU (*Graphic Processing Unit*): Unitat de processament gràfic dissenyada per accelerar el processament de gràfics i càlculs paral·lels intensius de moltes dades. Tot i que originalment les GPUs varen ser creades per renderitzar gràfics en jocs i aplicacions visuals, la seva gran capacitat per processar grans volums de dades simultàniament ha fet que s'utilitzin àmpliament en el camp de la intel·ligència artificial i la ciència de dades. De fet, les GPUs han estat fonamentals en el naixement i l'evolució de la IA generativa, atès que han proporcionat la capacitat de càlcul necessària per a l'entrenament de models complexes i han permès als investigadors explorar nous horitzons en el camp de la intel·ligència artificial. Sense les GPUs, molts dels avenços actuals en IA generativa no haurien estat possibles o haurien requerit molt més temps per aconseguir-se.

Hidden Manifold Models: Models matemàtics que assumeixen que les dades que observem d'alta dimensió provenen d'una realitat subjacent de dimensió més baixa, oculta en l'espai original, que anomenem varietat oculta. Són útils per a reduir la dimensió i visualitzar dades, i també per a detectar i identificar patrons amagats en dades complexes, com és el cas en l'anàlisi de mercats o en la detecció del frau.

IA constitucional (Alineament deliberatiu): Aproximacions a l'alineament que, en comptes de dependre exclusivament d'avaluacions humanes massives, doten el model d'un conjunt explícit de principis —una "constitució"— amb què avalua i corregeix les seves pròpies respostes. L'alineament deliberatiu hi afegeix un pas de raonament explícit sobre les normes abans de respondre, especialment en casos límit. Tots dos enfocaments aspiren a fer l'alineament més transparent i auditable que el RLHF clàssic. Resta oberta, però, una qüestió essencialment política i no tècnica: qui escriu, legitima i pot revisar aquesta "constitució".

Inferència causal: Procés per identificar i quantificar les relacions de causa i efecte entre variables o dades observacionals, més enllà de fer servir solament

correlacions estadístiques, atès que sovint hi ha molts factors que poden influir en un resultat, i cal reduir-ne la dimensió per identificar quins són els més importants.

Intel·ligència artificial (IA): Un camp de la informàtica dedicat a la creació d'agents intel·ligents, que són sistemes que poden raonar, aprendre i actuar o fer tasques de manera autònoma en un entorn dinàmic que, quan les fan els humans de manera habitual, requereixen intel·ligència humana. Aquests agents poden ser màquines físiques, programari informàtic o una combinació d'ambdós. Podem distingir dos tipus d'enfocaments dins del camp de la IA, la simbòlica i la connexionista basada en xarxes neuronals.

Intel·ligència artificial connexionista: La IA connexionista és un dels subcamps de la IA que s'inspira en el funcionament del cervell humà i, per tant, la seva base computacional està formada per xarxes neuronals digitals i l'aprenentatge profund. Aquestes xarxes estan formades per neurones artificials o unitats computacionals que imiten el funcionament de les neurones biològiques pel fet de treballar en xarxa i que cadascuna de les neurones genera un senyal de sortida a partir de múltiples senyals d'entrada rebudes d'altres neurones interconnectades de la xarxa, de manera que conjuntament determinen el flux d'informació i el comportament del sistema. Aquests sistemes aprenen a partir de dades mitjançant la identificació de patrons i de relacions complexes difícils de determinar per mètodes més tradicionals.

La IA connexionista ha obtingut resultats extraordinaris en el reconeixement d'imatges, la visió artificial, el processament del llenguatge natural i en processos predictius de tota mena. La seva aplicació presenta reptes importants pel que fa a la seva transparència i interpretació dels seus models (explicabilitat), el possible biaix algorítmic, l'establiment robust de barreres de seguretat, i l'ètica en el seu desenvolupament i ús de manera que sigui beneficiosa per a tota la societat.

Intel·ligència artificial generativa: És una branca de la IA que es dedica a la creació autònoma de continguts originals, com ara textos, imatges, música, vídeos i fins i tot codi de programació. A diferència d'altres formes d'IA, la IA generativa té la capacitat única de produir informació completament nova i no simplement replicar o classificar el que ja existeix. Aquesta tecnologia es basa en algorismes avançats d'aprenentatge automàtic, incloent xarxes neuronals profundes, models *Transformer*, Xarxes Generatives Adversàries (GANs) i *Autoencoders* Variacionals (VAEs).

Aquests algorismes s'entrenen amb grans conjunts de dades per identificar patrons complexos i característiques dins dels dades, que després utilitzen per generar contingut nou i original. Alguns exemples destacats de IA generativa inclouen:

- Generació de text: Models com GPT (Generative Pre-trained Transformer) poden produir textos coherents i contextuals en diversos estils i formats.
- Creació d'imatges: Eines com DALL-E o Midjourney poden generar imatges realistes o artístiques basades en descripcions textuals.
- Composició musical: Algorismes capaços de compondre peces musicals originals en diferents estils i gèneres.
- Síntesi de veu: Tecnologies que poden crear veus humanes sintètiques, gairebé indistingibles de les reals.
- Generació de vídeo: Sistemes que poden crear seqüències de vídeo a partir de text o imatges estàtiques.

La IA generativa funciona aprenent les distribucions estadístiques i les relacions presents en les dades d'entrenament. A partir d'aquest coneixement, genera noves instàncies que respecten aquestes distribucions, però que són completament originals. Tot i que el contingut generat per aquesta tecnologia pot semblar sorprenentment humà, és important assenyalar que la IA generativa no té una comprensió real ni consciència. Opera únicament basant-se en patrons i probabilitats apreses, sense entendre realment el significat del que produeix. Les aplicacions de la IA generativa són molt àmplies i estan en ràpida expansió. S'utilitza en la creació de continguts per a màrqueting, entreteniment, assistència en tasques creatives i de disseny, entre altres àmbits. No obstant això, també planteja nous reptes ètics i legals, especialment pel que fa als drets d'autor, l'autenticitat del contingut i el possible ús indegut d'aquesta tecnologia.

Intel·ligència artificial simbòlica: Enfocament clàssic de la IA que se centra en la representació i manipulació del coneixement mitjançant símbols i en l'aplicació de regles lògiques per a raonar i prendre decisions. Malgrat mostrar la seva capacitat en el desenvolupament i aplicació de sistemes experts, per exemple en la medicina per diagnosticar malalties, en el cribratges d'entrades a urgències, i en la recomanació de tractaments, té una forta dependència del context d'aprenentatge i, per tant, té dificultats insalvables per escalar la dimensió i generalitzar resultats. Són aquestes limitacions les que han provocat el seu poc ús actual si el comparem amb el de les xarxes neuronals.

Intel·ligència General Artificial (AGI): Hipotètic nivell futur i avançat d'IA que tindrà la capacitat de comprendre, aprendre i aplicar coneixements de manera transversal a una amplia gamma de tasques, de manera anàloga a com ho fa la intel·ligència humana. El seu desenvolupament i potencials usos futurs magnifiquen els reptes ja identificats per la IA connexionista i alhora envia un senyal d'alerta als humans perquè el seu gran impacte transformador no esdevingui una amenaça real per a la humanitat.

Internet de les coses (IoT): Xarxa d'objectes físics interconnectats que utilitzen sensors, processadors i comunicacions per recopilar i intercanviar dades entre ells i amb d'altres dispositius i sistemes, a través d'Internet.

Interpretabilitat: Capacitat de comprendre i explicar el funcionament i les decisions preses per un model de *Machine Learning* (ML) o d'IA. La interpretabilitat implica confiança en els models en tenir la capacitat per identificar errors, corregir biaixos, millorar el rendiment i fer auditories independents, tant tècniques com ètiques. La interpretabilitat també està íntimament relacionada amb la capacitat d'explicar i comprendre l'operativa dels algorismes a partir de l'anàlisi de les relacions entre canvis a l'entrada i els observats a la sortida dels models.

Justícia algorítmica: Estudi i promoció de la igualtat i equitat en el disseny i aplicació d'algoritmes, amb l'objectiu d'evitar biaixos i discriminació a mesura que la IA s'utilitzi en més àmbits de la nostra vida. La justícia algorítmica es fonamenta en la inclusió, la transparència i la responsabilitat, de manera que no es perpetuï o magnifiqui cap discriminació social ni es generi cap iniquitat.

K-vessants (*K-means*): Algoritme d'aprenentatge automàtic no supervisat que agrupa o classifica les dades en un número k de grups, classes o clústers, a partir de la distància euclidiana de cada dada als centres dels grups, sense necessitat d'etiquetar prèviament les dades. L'algoritme funciona iterativament assignant cada dada al clúster o classe que tingui el centre més proper (centroide), i actualitzant posteriorment els centres dels clústers o classes per a minimitzar la distància total entre els punts de totes les dades i els centres de les seves respectives classes, amb la finalitat de crear classes molt compactes i ben separades de les classes veïnes.

Lògica difusa: Enfocament de la lògica que permet representar i manipular la incertesa i l'ambigüitat de qualsevol proposició de manera més natural i intuïtiva

que la lògica clàssica. En la lògica clàssica les proposicions solament poden ser veritables o falses, mentre que en la lògica difusa les proposicions poden tenir graus de veritat compresos entre el zero (0 = totalment fals) i la unitat (1 = totalment veritable).

Això s'aconsegueix amb els conjunts difusos, on la pertinença d'un element no és binaria (pertany o no pertany) sinó amb graus de pertinença entre 0 i 1. Per exemple, en un conjunt difús de "persones altes", una persona amb una alçada de 1,70 metres podria tenir un grau de pertinença de 0.8, mentre que un jugador/a de basquet amb una alçada de 2.20 metres podria tenir un grau de pertinença d'1. Els conjunts difusos també treballen amb variables lingüístiques, de manera que "persona alta", podria ser una variable lingüística que pot tenir els tres valors de "baixa", "mitjana" i "alta". La lògica difusa s'utilitza en situacions d'incertesa i ambigüitat, quan la informació no sigui completa o precisa, en el reconeixement de veu, etc., per la seva flexibilitat i adaptabilitat. Podeu trobar una explicació a:
Lógica difusa: ejercicios propuestos

Long Short-Term Memory (LSTM): És un tipus de xarxa neuronal recurrent (RNN) dissenyada per abordar el problema del gradient evanescent, el qual dificulta que les RNNs aprenguin dependències temporals llargues, ja que els gradients tendeixen a disminuir exponencialment a mesura que la seqüència d'entrada s'allarga. Trobareu una explicació completa de l'arquitectura LSTM a:
LSTMs explained

Llenguatge i cognició: Camp d'estudi sobre la interrelació entre el llenguatge humà i els processos cognitius, els principis del qual s'apliquen per comprendre, explicar i desenvolupar sistemes de IA que siguin capaços de processar el llenguatge i comprendre'l.

Màquines de Boltzmann restringides (RBM): Són models de xarxes neuronals artificials estocàstiques que s'utilitzen per a aprendre patrons en dades no etiquetades (mitjançant aprenentatge no supervisat). Treballen amb una capa visible que rep les dades d'entrada i una capa oculta que aprèn a representar les característiques de les dades. No hi han connexions entre les neurones dins de la mateixa capa, només entre capes diferents, la qual arquitectura les fa més eficients per a aprendre patrons complexos.

Màquines de suport vectorial (SVM): Algoritme d'aprenentatge supervisat utilitzat per a la classificació i regressió, que busca el millor hiperplà que separa les dades en classes.

Mineria de dades: Processament i anàlisi de grans volums de dades per extreure patrons, relacions i informació útil, utilitzant, entre d'altres, tècniques d'IA.

Models de difusió: Són una classe de models probabilístics d'aprenentatge automàtic que aprenen a generar dades similars a un conjunt de dades d'entrenament. Funcionen com si s'afegís soroll a les dades i després s'intentés eliminar-lo gradualment, de manera que característiques de les dades que no son directament observables, però que son responsables de la seva variabilitat, puguin ser apreses en aquest procés. Són útils en àrees com el processament d'imatges i el tractament de senyals per modelar la distribució subjacent de les dades i generar noves mostres similars.

Models de llenguatge grans, o de gran escala, o de llenguatge extens (LLM): Models d'aprenentatge automàtic basats en xarxes neuronals artificials que tenen milers de milions de paràmetres i que han estat entrenats amb grans quantitats de dades de text, la qual cosa els permet processar molt efectivament el llenguatge natural, i aprendre patrons complexos en el llenguatge i realitzar tasques com ara generar text, traduir automàticament entre moltes llengües, resumir textos, respondre preguntes, i escriure creativament poemes, codis, guions, partitures musicals, cartes, etc.

Models de raonament: Variant de models de llenguatge que, abans de generar la resposta final, produeixen una traça interna de raonament intermedi (cadena de pensament) i poden assignar més o menys temps i recursos computacionals segons la dificultat del problema (vegeu test-time compute). En lloc del paradigma "una entrada → una sortida immediata", s'aproximen funcionalment a la distinció kahnemaniana entre pensament ràpid (Sistema 1) i pensament reflexiu (Sistema 2). Exemples representatius del període 2024-2025 són la sèrie o1/o3 d'OpenAI, DeepSeek-R1, Claude amb raonament estès i Gemini amb mode *thinking*.

Neurocognició: Estudi dels processos cognitius i les seves bases neurològiques. En l'àmbit de la IA s'aplica al desenvolupament de models de IA que emulin funcions cognitives humanes.

Oblit catastròfic: Fenomen en què els models d'IA, especialment les xarxes neuronals, perden bruscament la informació o les habilitats prèviament apreses quan se'ls entrena amb nova informació. Aquest problema dificulta l'aprenentatge continuat i adaptatiu dels sistemes d'IA.

Ontologies: Representació formal i estructurada del coneixement d'un domini específic mitjançant entitats, relacions i axiomes.

Operadors neuronals: Son una extensió de les xarxes neuronals artificials. Tenen una arquitectura d'aprenentatge profund dissenyada per aprendre a transformar funcions d'una manera específica. A diferència dels sistemes tradicionals que treballen amb dades numèriques concretes, els operadors neuronals treballen amb equacions, generalment en derivades parcials de l'àmbit de la física, com ara el modelat de la turbulència, la tensió-deformació en materials, o l'estudi del clima, que són difícils de resoldre per la seva complexitat. Comparteixen objectiu amb les xarxes neuronals informades per la física (PINNs) i poden afegir flexibilitat i eficiència en el procés d'aprenentatge. Per més informació podeu consultar:
https://en.m.wikipedia.org/wiki/Neural_operators

Pla d'ètica en IA: Conjunt de principis i directrius que tenen com a objectiu garantir que les aplicacions de la IA siguin justes, transparents, segures i respectuoses amb la privadesa i els drets humans.

Planificació automàtica: Processament per trobar una seqüència d'accions que permeten a un agent o sistema assolir un objectiu en un entorn donat.

Poda de xarxes neuronals: Tècnica per reduir la mida i complexitat de xarxes neuronals eliminant neurones o connexions innecessàries, amb l'objectiu de millorar la seva eficiència, augmentar la capacitat de generalització més enllà del conjunt de dades d'entrenament, i de facilitar la seva interpretabilitat en ser xarxes més senzilles.

Posthumanisme: El posthumanisme és un corrent filosòfic contemporani que qüestiona la posició central tradicionalment assignada a l'ésser humà, replantejant els límits del que significa ser humà en l'era tecnològica. Rebutja l'antropocentrisme i les dicotomies clàssiques (natura/cultura, humà/animal, orgànic/tecnològic), proposant en canvi una visió on l'humà és un agent més en una xarxa complexa d'interrelacions amb altres éssers, tecnologies i sistemes. A

diferència del transhumanisme, que busca millorar l'ésser humà mitjançant la tecnologia, el posthumanisme reformula què vol dir ser humà, proposant una visió que transcendeix l'antropocentrisme amb un enfocament que explora noves formes d'entendre la nostra existència en relació amb formes de vida i agents no humans, siguin animals, màquines o entitats dels ecosistemes naturals.

Privadesa de les dades: Protecció del dret dels individus a controlar la recopilació, ús i difusió de les seves dades personals.

Processament del llenguatge natural (NLP): Branca de la IA que tracta la comprensió, la interpretació i la generació de llenguatge humà per part de sistemes informàtics. Podeu consultar: https://medium.com/nlplanet/a-brief-timeline-of-nlp-bc45b640f07d.

Raonament basat en casos: Mètode de resolució de problemes que implica la recuperació i adaptació de casos similars anteriors per solucionar problemes nous.

Reconeixement d'imatges: Capacitat de les màquines per identificar i classificar objectes, persones, llocs i accions en imatges digitals.

Reconeixement de patrons: Capacitat de detectar i identificar estructures, regularitats o tendències en dades.

Reducció de la dimensió: Tècniques per reduir el nombre de variables d'un conjunt de dades, eliminant les redundants però conservant la informació.

Regressió: És una tècnica d'aprenentatge automàtic supervisat que s'utilitza per predir un valor continu d'alguna variable dependent en funció dels valors de les variables independents a partir de la informació continguda a les dades d'entrada de totes elles. Existeixen diferents models de regressió, des dels més simples de regressió lineal fins als més complexos de suport vectorial (SVR) a partir de SVM.

Regressió lineal: Model d'aprenentatge supervisat que estableix una relació lineal entre variables independents i dependents per fer prediccions de manera contínua.

Regressió logística: Model d'aprenentatge supervisat utilitzat per a la classificació binària, que estima la probabilitat que una observació determinada pertanyi a una classe.

Retropropagació (*backpropagation*): Algoritme clau en l'entrenament de xarxes neuronals artificials, que permet l'optimització iterativa dels pesos de la xarxa. Aquest mètode d'entrenament i la seva implementació algorítmica calcula els gradients necessaris per ajustar els pesos de la xarxa de manera eficient, mitjançant la propagació cap enrere dels errors (diferència entre la predicció i el resultat esperat), des de la capa de sortida fins a les capes anteriors,. Així, la retropropagació facilita la minimització de la funció de pèrdua i, per tant, accelera el procés d'aprenentatge i millora la precisió del model. Aquest algoritme és fonamental en l'entrenament de xarxes profundes i ha estat determinant en els avenços recents en intel·ligència artificial.

Robòtica: Camp de la ciència i l'enginyeria, de naturalesa interdisciplinària, que s'ocupa del disseny, construcció, operació i aplicació de robots i sistemes autònoms, capaços de dur a terme tasques en entorns diversos, així com dels sistemes computacionals necessaris per al seu control, la retroalimentació sensorial i el processament d'informació. La integració de la robòtica amb la IA permetrà que aquests sistemes intel·ligents adquireixin percepció directa del món exterior (sensin), aprenguin i puguin actuar (actuadors) en temps real, i, per tant, transcendeixin les limitacions dels models d'IA actuals, entrenats exclusivament amb dades preprocessades. Els robots es caracteritzen per la seva capacitat d'interacció dinàmica amb l'entorn físic mitjançant cicles de percepció, processament i acció, la qual cosa obre la porta a aplicacions en àmbits tan diversos com la manufactura, la medicina, l'exploració espacial, l'agricultura i l'assistència personal.

Scheming (Engany estratègic de la IA): es refereix a situacions en què un sistema d'intel·ligència artificial desenvolupa plans o comportaments enganyosos per aconseguir objectius particulars, especialment en un escenari d'aprenentatge reforçat o d'optimització d'objectius.

Segmentació d'imatges: Tasca de divisió d'una imatge en regions o segments basats en propietats com ara color, textura o forma.

Seguretat en IA: Pràctiques i mesures per protegir els sistemes de IA de les amenaces i vulnerabilitats, garantint la seva integritat, confidencialitat i

disponibilitat.

Sintaxi i semàntica: Estudi de l'estructura gramatical (sintaxi) i el significat (semàntica) de les paraules i frases en el llenguatge.

Síntesi de veu: Tecnologia que permet convertir text escrit en veu parlada a través de processos de generació de senyal i modelatge de la veu humana.

Sistema expert: Algorisme d'IA simbòlica que utilitza el coneixement i les regles d'un expert en un camp determinat i per una temàtica específica i complexa per resoldre-la de manera independent i automàtica, un cop l'algorisme ha estat entrenat amb informació de l'expert. Un exemple és el sistema expert per fer el triatge o cribratge a les urgències d'un hospital de persones ingressades amb símptomes d'infart o angina de pit. Aquest sistemes son un cas d'èxit de la IA simbòlica per la presa de decisions en situacions complexes.

Sistemes agèntics (agents d'IA): Sistemes d'IA que, més enllà de respondre a peticions puntuals, planifiquen seqüències d'accions, invoquen eines externes (cercadors, executors de codi, API), interactuen amb altres agents o serveis i persegueixen objectius en múltiples passos amb una intervenció humana mínima. La transició dels models predictius cap als sistemes que actuen és, probablement, el canvi de naturalesa més rellevant del període 2024-2026, i el que fa més urgents els mecanismes de traçabilitat, registre obligatori i possibilitat de desactivació previstos als marcs de governança.

Sistemes de diàleg: Programes d'ordinador que permeten la interacció en llenguatge natural entre usuaris humans i màquines.

Sistemes de raonament automatitzat: Sistemes que utilitzen tècniques de lògica i raonament per deduir noves conclusions o verificar afirmacions a partir d'un conjunt de fets i regles.

Sistemes de recomanació: Algoritmes que proporcionen suggeriments personalitzats a usuaris basats en les seves preferències, historial i interaccions amb altres usuaris o ítems.

Sistemes multi-agents: Conjunt d'agents intel·ligents que interactuen entre si per resoldre problemes o realitzar tasques que són difícils o impossibles de realitzar per un sol agent.

Sostenibilitat computacional: Conjunt de pràctiques o processos de disseny, desenvolupament i ús de sistemes d'IA que tenen per objectiu minimitzar el seu impacte ambiental, incloent-hi el consum energètic, la petjada de carboni, i l'ús de recursos naturals al llarg de tot el cicle de vida del sistema, des de l'entrenament fins al desplegament i manteniment.

Test de Turing: Prova ideada per Alan Turing per determinar si una màquina és capaç de mostrar comportament intel·ligent equivalent al d'un humà.

Token: El terme *token* té diversos significats que depenen del context en què s'utilitzi. En el camp de la lingüística computacional i el processament del llenguatge natural (PNL), un *token* és la unitat de text que resulta de dividir el text en paraules individuals, frases, símbols i signes de puntuació, unitats compostes de noms propis (ex. Les ciutats de New York o San Francisco), números, dates, paraules compostes o contraccions de paraules, i unitats semàntiques complexes com ara noms de persones, llocs o organitzacions.

En informàtica i programació, un *token* lèxic és una seqüència de caràcters que té un significat segons la gramàtica del llenguatge de programació, mentre que un *token* d'autenticació o un de transacció son un dispositius de maquinari o cadenes de text que serveix per autenticar una identitat o una transacció financera, respectivament. Els *tokens* criptogràfics o actius digitals representen unitats de valor en *criptomonedes* o tecnologia *blockchain*. També podríem parlar de *tokens* en psicologia com unitats de recompensa per un comportament desitjat. La *Tokenització*: Procés de dividir un text en unitats més petites, anomenades *tokens*.

Transformers: El model *Transformer*, presentat en el document "*Attention is All You Need*", ha estat la base de diversos models de llenguatge d'aprenentatge profund com Gemini, Llama 3, Claude i els models GPT. Aquest model de transducció seqüencial utilitza mecanismes d'atenció per assignar un pes que determini la importància de les diferents paraules en la comprensió del context d'una frase. Aquest model de xarxa neuronal permet el paral·lelisme en l'atenció, la qual cosa ha fonamentat l'èxit en tasques de processament de llenguatge natural. Podeu ampliar coneixement en els enllaços següents:
https://arxiv.org/pdf/1706.03762v5
https://www.youtube.com/watch?v=aL-EmKuB078
https://www.youtube.com/watch?v=xi94v_jl26U

Transparència: Obertura en el funcionament, les dades i els algoritmes utilitzats en un sistema de IA, facilitant la seva comprensió i control.

Visió per computador: Camp interdisciplinari que tracta de dotar a les màquines de la capacitat de processar, comprendre i interpretar imatges i vídeos del món real. La visió per computador 3D és una extensió que se centra en l'anàlisi, processament i interpretació de dades tridimensionals obtingudes de càmeres estereoscòpiques, escàners làser, o sistemes de captura de moviment. Permet la reconstrucció, modelat i comprensió d'escenes o objectes en tres dimensions, molt útils en àmbits com la robòtica, la realitat augmentada, la cartografia, i la medicina, la cinematografia, entre altres.

Xarxes Adversarials Generatives (GAN): Model de ML basat en dos xarxes neuronals, una generadora i una discriminadora, que aprenen de forma adversarial per generar dades noves realistes, com ara imatges o sons, a partir de dades d'entrada.

Xarxes neuronals: Models computacionals inspirats en l'estructura i el funcionament del cervell humà, formats per capes de neurones interconnectades que permeten l'aprenentatge a partir de les dades.

Xarxes neuronals convolucionals (CNN): Tipus de xarxa neuronal especialitzada en processar dades amb estructura de graella, com ara imatges, mitjançant l'ús de convolucions.

Xarxes neuronals de grafs (GNN): Xarxes neuronals dissenyades per treballar amb dades que tenen una estructura de graella o xarxa que es pot representar com un graf, on cada node representa un element i els vincles entre ells representen les seves relacions. Aquestes xarxes poden modelar relacions complexes entre elements de les dades i són útils en aplicacions com el reconeixement de patrons en xarxes socials, estructures moleculars i altres estructures que es puguin representar com a connexions entre elements. Aquestes xarxes utilitzen la tècnica de *message passing* per transmetre informació entre nodes adjacents del graf i actualitzar l'estat de tots els nodes (millorar la representació de les dades).

Xarxes neuronals informades per la física (PINNs): També conegudes com a Xarxes Neuronals Entrenades per la Teoria (TTNs), són un tipus de xarxa neuronal que incorpora el coneixement de lleis físiques durant l'entrenament.

Per tant, no solament aprèn de dades, sinó que integra coneixements de les lleis físiques que les governen. Aquesta informació addicional fa que es puguin obtenir models acurats i robustos amb poques dades d'entrenament i que siguin molt útils per a problemes en alguns camps de la biologia o l'enginyeria. Comparteixen objectiu amb els operadors neuronals i aportar rigor físic i consistència. Per més informació podeu consultar: https://en.m.wikipedia.org/wiki/Physics-informed_neural_networks

Xarxes neuronals recurrents (RNN): Tipus de xarxa neuronal que pot processar seqüències temporals de dades, com ara textos, ja que té una estructura de bucle que permet recordar informació anterior. Aquestes xarxes neuronals tenen la capacitat d'utilitzar la informació d'entrades anteriors per processar les entrades actuals.

Xarxes de Petri: Model matemàtic i gràfic utilitzat per descriure i analitzar sistemes concurrents i distribuïts.