Glossari alfabètic d'intel·ligència artificial

AGI (INTEL·LIGÈNCIA ARTIFICIAL GENERAL): concepte teòric que es refereix a una IA capaç d'igualar o superar el rendiment cognitiu humà en pràcticament qualsevol tasca, amb la capacitat de raonar i aprendre a través de múltiples dominis de manera autònoma.

AJUST FI (*FINE-TUNING*): procés de prendre un model d'IA ja entrenat en un conjunt de dades general i realitzar un entrenament addicional amb dades específiques per a optimitzar el seu rendiment en una tasca o domini particular, com la medicina, el dret, l'atenció al client o l'educació. Sol ser més eficient que crear un model nou.

AL·LUCINACIÓ (*HALLUCINATION*): ocorre quan un model generatiu produeix informació que sona molt convincent, però és fàcticament falsa, inventada o confusa, pel fet que el model prediu la paraula més probable sense consultar una base de dades fidedigna. Les al·lucinacions també poden deure's a biaixos en els conjunts de dades d'entrenament o a la falta d'accés del model a informació actualitzada.

ALGORISME (*ALGORITHM*): conjunt de regles, fórmules o instruccions emprades per a processar informació i obtenir un resultat. En l'àmbit de la IA, l'algorisme és una indicació per al sistema que permet aprendre d'exemples, detectar patrons o classificar informació. És un procediment ordenat per a resoldre un problema.

ANI (IA LIMITADA O FEBLE): models d'IA dissenyats i entrenats per a fer una tasca específica i molt concreta que no poden adaptar-se a altres dominis fora de la seua programació. Són assistents virtuals, traductors automàtics, reconeixement d'imatges, etcètera.

ANOTACIÓ DE DADES: procés d'afegir etiquetes, marques o descripcions específiques a les dades sense processar, com a imatges o textos, perquè els algorismes d'IA els entenguen i puguen reconèixer objectes, accions o conceptes durant el seu entrenament.

APRENENTATGE AUTOMÀTIC (*MACHINE LEARNING - ML*): branca de la IA que permet als sistemes aprendre i adaptar-se sense necessitat de ser programats amb instruccions explícites per a cada tasca. Utilitza algorismes per a aprendre de les dades i millora la seua precisió de manera gradual.

APRENENTATGE NO SUPERVISAT: enfocament d'entrenament en què la màquina busca patrons, estructures o grups ocults per si mateixa dins de conjunts de dades

que no tenen etiquetes ni instruccions prèvies sobre el que ha de trobar. És útil quan hi ha moltes dades disponibles però no es disposa d'etiquetes ja preparades.

APRENENTATGE PER REFORÇ (*REINFORCEMENT LEARNING*): tipologia específica d'aprenentatge automàtic en la qual la màquina aprèn a partir dels errors i encerts comesos en una tasca concreta mitjançant un sistema de recompenses i castics que guien el seu comportament. La idea és semblant a aprendre per experiència: repetir el que funciona i evitar el que dona mal resultat.

APRENENTATGE PROFUND (*DEEP LEARNING*): subcamp de l'aprenentatge automàtic que utilitza xarxes neuronals artificials amb moltes capes per a modelar i entendre patrons complexos en grans conjunts de dades, imitant l'estructura del cervell humà. És molt útil quan la informació és difícil de descriure.

APRENENTATGE SEMISUPERVISAT: branca del *machine learning* que combina l'aprenentatge supervisat i no supervisat mitjançant l'ús de dades etiquetades i no etiquetades per a entrenar models, i és útil quan hi ha grans volums de dades sense etiquetar.

APRENENTATGE SUPERVISAT: model d'entrenament en què la màquina aprèn utilitzant dades que ja han sigut etiquetades amb la resposta correcta, la qual cosa permet al sistema comparar el seu resultat amb la solució real per a aprendre de forma guiada. Durant l'entrenament, compara la seua predicció amb la resposta i es corregeix d'una manera progressiva.

BIAIX ALGORÍTMIC (*ALGORITHMIC BIAS*): anomalia o prejudici en els resultats de la IA produït perquè les dades d'entrenament contenien prejudicis humans, històrics o per suposicions errònies durant el disseny de l'algorisme.

CAIXA NEGRA (*BLACK BOX*): sistemes, inclosos diversos algorismes d'aprenentatge automàtic, el funcionament dels quals és inaccessible o massa complex perquè l'usuari el comprenga fàcilment, fet que implica falta de transparència en sistemes d'IA complexos en què les operacions internes són tan intricades que resulta difícil per als humans, inclosos els creadors, explicar exactament per què l'algorisme ha pres una decisió específica.

CODI OBERT: model de desenvolupament en què el codi d'un sistema està disponible per a ser revisat, reutilitzat o modificat per altres persones. Afavoreix l'escrutini, la col·laboració i, en molts casos, una millora ràpida d'aquest. En IA, parlar de codi obert no sempre significa que tot estiga obert. El codi pot ser públic,

mentre que les dades, els pesos del model o unes certes condicions d'ús poden ser restringits.

DADES D'ENTRENAMENT (*TRAINING DATASETS*): conjunt de dades que s'utilitza per a ensenyar als models d'aprenentatge automàtic a identificar patrons i generar resultats. La qualitat, diversitat, actualitat i equilibri d'aquestes dades influeixen directament en el comportament del sistema. Si les dades són pobres o estan esbiaixades, el model tendirà a reflectir aquestes limitacions.

DEEPFAKES: imatges i vídeos alterats deliberadament perquè semblen realistes. Pot imitar la veu, aparença o els gestos d'una persona. Planteja riscos importants de frau, desinformació i suplantació.

DESINFORMACIÓ: creació i difusió deliberada d'informació falsa o manipulada que té com a objectiu influir, confondre o enganyar les persones, ja siga amb la finalitat de causar mal o amb finalitats polítiques, personals o econòmiques. No és un error sinó una acció orientada a produir una percepció de la realitat equivocada.

ENGINYERIA DE *PROMPTS* (*PROMPT ENGINEERING*): pràctica de dissenyar, refinar i optimitzar les instruccions d'entrada per a aconseguir que el model d'IA retorne el resultat més precís, útil i alineat amb els objectius de l'usuari.

FINESTRA DE CONTEXT (*CONTEXT WINDOW*): quantitat màxima d'informació o *tokens* que el model pot tenir en compte en una interacció determinada. Inclou el que la persona escriu i, segons el sistema, també part del que s'ha dit anteriorment. Com més gran n'és la finestra, més text o dades pot considerar el sistema abans de perdre detalls rellevants.

IA DE FRONTERA (*FRONTIER IA*): segons la Junta Assessora Científica de les Nacions Unides es refereix a models d'IA de propòsit general altament capaços que poden fer moltes tasques igualant o superant les capacitats dels models més avançats d'avui dia. Són els models més avançats.

IA DE PROPÒSIT GENERAL: sistema avançat d'IA capaç de realitzar eficaçment una sèrie de tasques distintes i adaptar-se a una àmplia gamma d'aplicacions i no a una sola aplicació. Pot resumir, redactar, traduir, classificar extraure informació. Cal no confondre-la amb la IA general, ja que la IA de propòsit general és més limitada.

IA EXPLICABLE (*xAI*): descriu la capacitat de presentar o explicar el procés de presa de decisions d'un sistema d'aprenentatge automàtic en termes que puguen ser entesos per humans per a garantir la transparència i la confiança.

IA GENERATIVA (*GENERATIVE AI*): tipus d'IA capaç de crear contingut original i nou, incloent-hi text, imatges, àudio, vídeo o altres mitjans, basant-se en patrons apresos a partir de grans volums de dades existents en resposta a les instruccions de l'usuari.

IA MULTIMODAL: IA que combina i analitza diferents entrades d'informació alhora, com a text, imatge, àudio o vídeo, combina diverses fonts per a produir la resposta més completa i generar resultats més sòlids. Permet a la IA, per exemple, analitzar una fotografia, llegir una pregunta sobre aquesta i respondre amb text.

IA RESPONSABLE: conjunt de principis que ajuden a guiar el disseny, desenvolupament, desplegament i ús de la IA, la qual cosa genera confiança en les solucions d'IA que tenen el potencial d'empoderar les organitzacions i les seues parts interessades. No se centra sols a fer que la tecnologia funcione sinó en els seus efectes en la societat.

INTERPRETABILITAT: capacitat de presentar o explicar el procés de presa de decisions d'un sistema d'aprenentatge automàtic en termes que puguen ser entesos per humans. També es denomina *transparència* o *explicabilitat*.

MECANISME D'ATENCIÓ (*ATTENTION MECHANISM*): tècnica que permet al model enfocar-se en les parts més rellevants d'un text o dada per a entendre el seu significat profund, i assignar importància a les paraules clau segons el context de la frase.

MODELS DE LLENGUATGE EXTENS (*LARGE LANGUAGE MODELS*, LLM): tipus de model fundacional entrenat amb quantitats massives de text per a comprendre i generar llenguatge humà natural, que és capaç de fer tasques complexes com ara traducció, resum i redacció creativa. Durant les fases d'entrenament, els grans models de llenguatge aprenen paràmetres a partir de factors com la grandària del model i els conjunts de dades d'entrenament. Els models de llenguatge extens utilitzen els paràmetres per a inferir nou contingut.

MODELS FUNDACIONALS (*FOUNDATION MODELS*): models d'aprenentatge automàtic massius entrenats amb conjunts de dades gegantesques per a adquirir coneixements generals que després poden adaptar-se a una gran varietat d'aplicacions i tasques específiques posteriors. Serveixen com a base o blocs de construcció per a crear aplicacions més especialitzades.

NLP (PROCESSAMENT DE LLENGUATGE NATURAL): camp de la IA que s'ocupa de la interacció entre les computadores i el llenguatge humà per a permetre a les

màquines entendre, interpretar i generar el llenguatge humà de manera natural i coherent.

PARÀMETRES: connexions internes, els ajustos numèrics, d'un model que canvien durant l'entrenament i determinen la seua capacitat per a captar matisos complexos i minimitzar l'error. El seu nombre sol ser un indicador de la potència i sofisticació del sistema d'IA.

PRESA DE DECISIONS AUTOMATITZADA: ús de la tecnologia, especialment algorismes d'intel·ligència artificial i aprenentatge automàtic, per a prendre decisions de manera autònoma. Implica processar dades, analitzar-les i avaluar opcions sense intervenció humana constant.

PRIVACITAT DIFERENCIAL: conjunt de tècniques matemàtiques utilitzades per a entrenar models assegurant que no es puguen extraure ni identificar dades personals específiques dels individus inclosos en el conjunt d'entrenament.

PROMPT: instrucció, pregunta o text d'entrada que l'usuari proporciona a una IA generativa per a iniciar una tasca o obtenir una resposta específica del model. El *prompt* pot ser molt breu o molt detallat. Com millor expresse l'objectiu, el context i el format esperat més probable és que la resposta siga útil.

RLHF (APRENENTATGE PER REFORÇ A PARTIR DE RETROALIMENTACIÓ HUMANA): tècnica de *machine learning* en què un model de recompensa s'entrena amb comentaris humans directes i després s'utilitza per a optimitzar el rendiment d'un agent d'intel·ligència artificial mitjançant l'aprenentatge per reforç.

SOBIRANIA TECNOLÒGICA: capacitat d'un país, institució o sector per a disposar de la infraestructura, el coneixement i el control sobre tecnologies clau per a desenvolupar IA. Això suposa no dependre per complet de proveïdors externs per a dades, computació, models o serveis crítics.

SUPERINTEL·LIGÈNCIA: sistema d'intel·ligència artificial (IA) hipotètic, basat en programari amb un abast intel·lectual que va més enllà de la intel·ligència humana. En el nivell més fonamental, aquesta IA superintel·ligent disposa de funcions cognitives d'avantguarda i capacitats de pensament molt desenvolupades, més avançades que les de qualsevol ésser humà.

TECNOLOGIA EDUCATIVA: conjunt d'eines digitals destinades a donar suport a l'ensenyament i l'aprenentatge. No tota tecnologia educativa utilitza IA, però moltes

eines actuals la incorporen en les plataformes de formació, sistemes de tutoria, continguts interactius, entre d'altres.

TEMPERATURA: paràmetre de configuració que controla l'aleatorietat i creativitat del model. Una temperatura baixa dona respostes més predictibles i una d'alta genera resultats més variats i originals.

TOKENITZACIÓ: procés de dividir el text en unitats més petites anomenades *tokens*, que poden ser paraules, síl·labes o caràcters, perquè els models d'IA puguen processar el llenguatge de manera numèrica i matemàtica. 1.000 *tokens* equivalen a unes 750 paraules.

TRANSFORMADORS (*TRANSFORMERS*): arquitectura de xarxa neuronal que permet processar seqüències de dades de manera paral·lela i analitzar frases completes alhora, la qual cosa facilita una comprensió molt superior del context en comparació amb models anteriors.

VISIÓ ARTIFICIAL (*COMPUTER VISION*): subcamp de la intel·ligència artificial que dota les màquines de la capacitat de processar, analitzar i interpretar entrades visuals com ara imatges i vídeos. Utilitza *machine learning* per a ajudar els ordinadors i altres sistemes a obtenir informació significativa a partir de dades visuals.

XARXA GENERATIVA ADVERSARIAL (GAN): xarxes de desenvolupament recent en IA formades per dues subxarxes neuronals artificials: una xarxa generadora i una xarxa discriminadora. La xarxa generadora rep dades d'entrenament i produeix dades artificials basades en patrons. La xarxa discriminadora compara les dades generades artificialment amb les dades d'entrenament reals i transmet a la xarxa generadora on ha detectat diferències. El generador llavors altera els seus paràmetres. Amb el temps, la xarxa generadora aprèn a donar dades més realistes fins que la xarxa discriminadora no pot distingir què és artificial i què és real. L'objectiu és millorar i que el sistema arribe a produir resultats cada vegada més realistes.

XARXA NEURONAL ARTIFICIAL (*ARTIFICIAL NEURAL NETWORK*): arquitectura informàtica inspirada en el cervell biològic composta per unitats computacionals interconnectades anomenades neurones, organitzades per capes, que treballen conjuntament per a processar informació.

XARXA TEAM (EQUIP ROIG): pràctica de fer proves de seguretat en què els experts intenten trobar de manera deliberada vulnerabilitats, biaixos o comportaments

perillosos en un sistema d'IA per a corregir-los abans que es desplegue públicament. Consisteix a intentar que el sistema falle per a entendre millor els seus riscos.