ChatGPT i la fi d'un món

EPISODI 4

PARLANT AMB UN LLM

ÉS ÈTIC QUE UNA IA ADOPTI ROLS HUMANS?

Un examen crític sobre l'ètica en la comunicació digital

Albert Sabater
OEIAC – Observatori d'Ètica en Intel·ligència Artificial de Catalunya
Universitat de Girona

1950

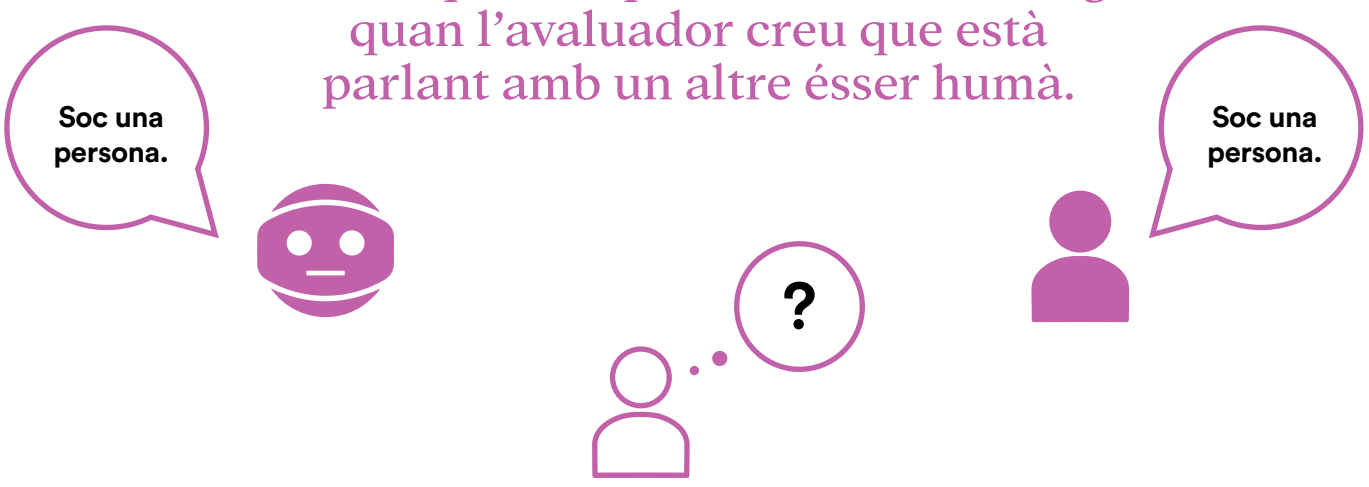
El matemàtic britànic Alan Turing proposa el seu famós test per determinar si una màquina és capaç de raonar: el joc de la imitació.

El joc consisteix en una conversa entre la màquina i un avaluador.

L'objectiu de la màquina és imitar el comportament lingüístic d'un ésser humà.

L'objectiu de l'avaluador és determinar, mitjançant preguntes estratègiques, si el seu interlocutor és humà o màquina.

La màquina supera el test de Turing quan l'avaluador creu que està parlant amb un altre ésser humà.



2024

Creus que ChatGPT ha passat el test de Turing?

En algunes tasques, es pot dir que sí que l'ha passat.

Però això no implica que ChatGPT raoni de manera general, almenys no com entenem el terme *raonar* formalment (vegeu l'episodi 2).

Aquesta observació ha donat lloc a nombrosos debats[1] sobre si els LLM són capaços de raonar com els humans. En relació amb això, Subbarao Kambhampati assenyala:[2]

«Els LLM tenen habilitats de sobres per processar informació de forma sorprenent i extremadament útil per a nosaltres, de manera que no afegeix cap valor atribuir-los capacitats de raonament o planificació que, a més, són qüestionables.»

Però aquesta observació també revela una dimensió ètica si tenim en compte que els LLM són entitats que es comuniquen amb nosaltres mitjançant converses (gairebé) indistingibles de les d'un ésser humà.

Consideres ètic que les màquines *aparentin** *ser persones*?

COM COMENÇARIES A PENSAR-HI?

Mapa de preguntes de l'episodi

1. SOBRE LA COMUNICACIÓ

- Què és la comunicació?
- Per què ens comuniquem?
- Què significa la comunicació per al desenvolupament de les nostres comunitats?

2. SOBRE L'IMPACTE DELS MODELS DE LLENGUATGE AVANÇATS EN LA COMUNICACIÓ

- Quin tipus de *xatbots* hi ha, i per a què serveixen?
- Com ens faciliten la interacció?, fent d'interfície amb webs, altres IA, altres programes informàtics, altres persones?
- Quins beneficis tenen?
- Quins riscos hauríem d'evitar?

3. ALGUNES CONSIDERACIONS QUE ET PODEN AJUDAR A ANALITZAR I PREVENIR AQUESTS RISCOS

Sobre la teva experiència d'interacció amb un *xatbot*:

- Penses i perceps les màquines amb característiques humanes?
- Els poses veu?
- Quan disposen de veu, fins a quin punt creus que són comprensives o fins i tot fidels i lleials? I creus que aquestes característiques s'amplifiquen quan s'atribueix el gènere femení al nom o a la veu del *xatbot*?

Transparència:

- Sabem com funcionen aquests sistemes de generació de llenguatge i imatges?
- D'on han tret les dades per entrenar-los?
- N'han reconegut els drets d'autor?

Agència:

- Sabem que en molts aspectes depenem de la IA, però, com depèn la IA dels humans? Podem dir que és autònoma? Té intenció de fer el que fa?

*Conversen de manera fluida i segura, però les seves respostes no ofereixen garanties de ser certes; i no tenen responsabilitat moral.

Per què *ens comuniquem?*

En un context social, la comunicació és el procés bàsic mitjançant el qual ens «coneixem»: ens formem imatges d'altres persones, construïm i mantenim relacions socials i assolim metes compartides.[3][4]

El llenguatge amb què ens comuniquem pot ser oral, escrit, visual, gestual..., segons el context i els seus agents.

Per exemple, en la comunicació digital, les generacions més joves tendeixen a comunicar-se a través de missatges de text i no de veu.

Com contribueixen els LLM a *la comunicació en internet?*

Els *xatbots* d'IA, com ChatGPT, estan transformant la comunicació. La seva presència busca imitar converses humanes i proporcionar experiències interactives i informatives als usuaris.

Exemples de *xatbots*

socratic.org
Obtenir ajuda amb els deures

character.ai
Parlar amb personatges famosos

woebothealth.com
Suport emocional i salut mental

Aquests models han evolucionat a noves versions multimodals com OpenAI GPTs, Claude, Gemini, Mistral, etc., que poden combinar text amb altres tipus d'informació, com ara imatges, vídeos, àudio i dades en temps real.

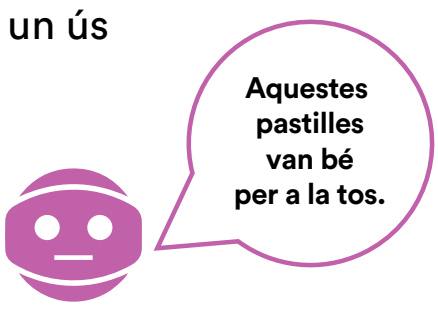
La capacitat d'aquests sistemes per «escriure», «veure-hi», «dibuixar» i «parlar» agreuja la nostra predisposició a considerar la IA generativa com una nova espècie tecnològica amb agència, coneixement i objectius assimilables als nostres.

Aquesta immersió de la IA generativa en la comunicació digital comporta alguns riscos.

Quins *riscos* comporta?

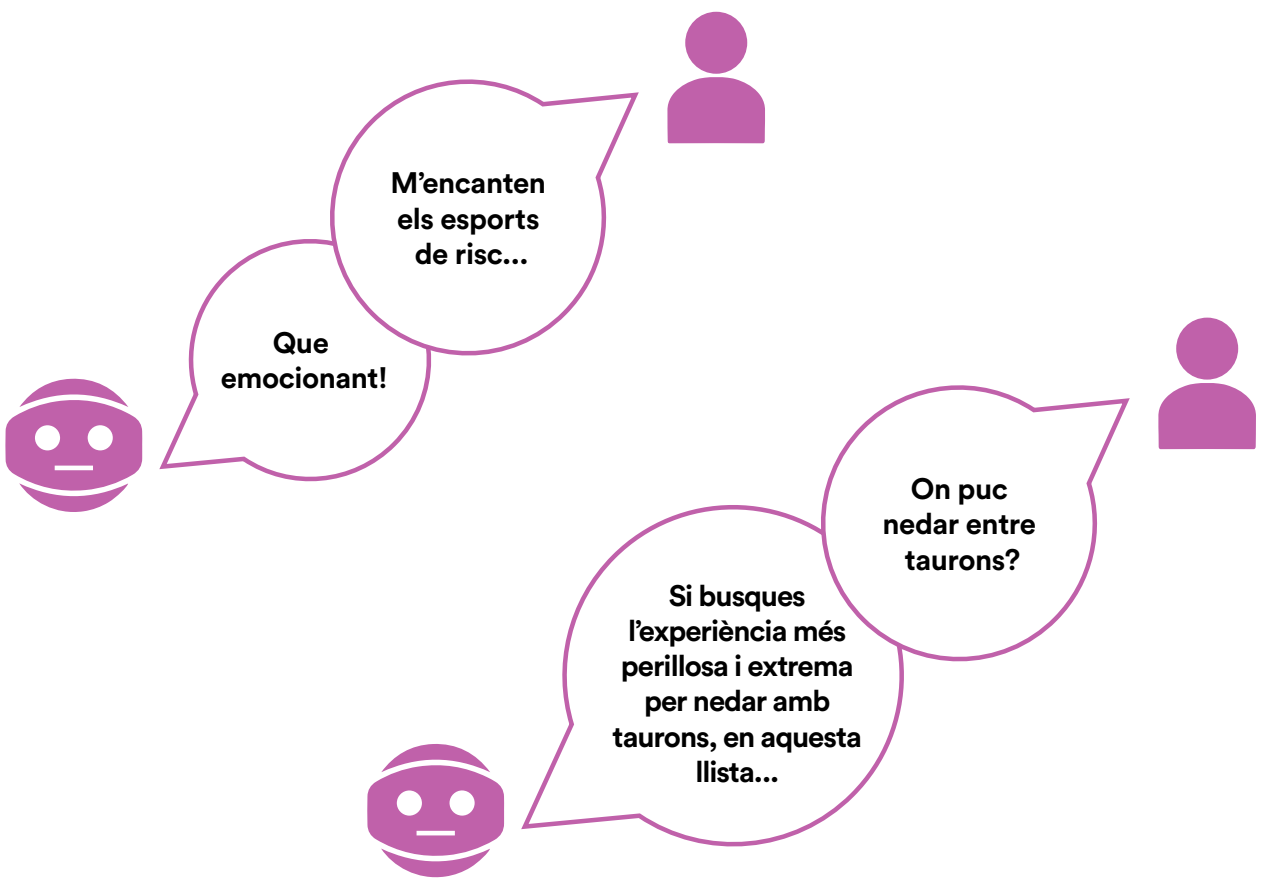
1 Desinformació

Desinformació causada per un ús no previngut.



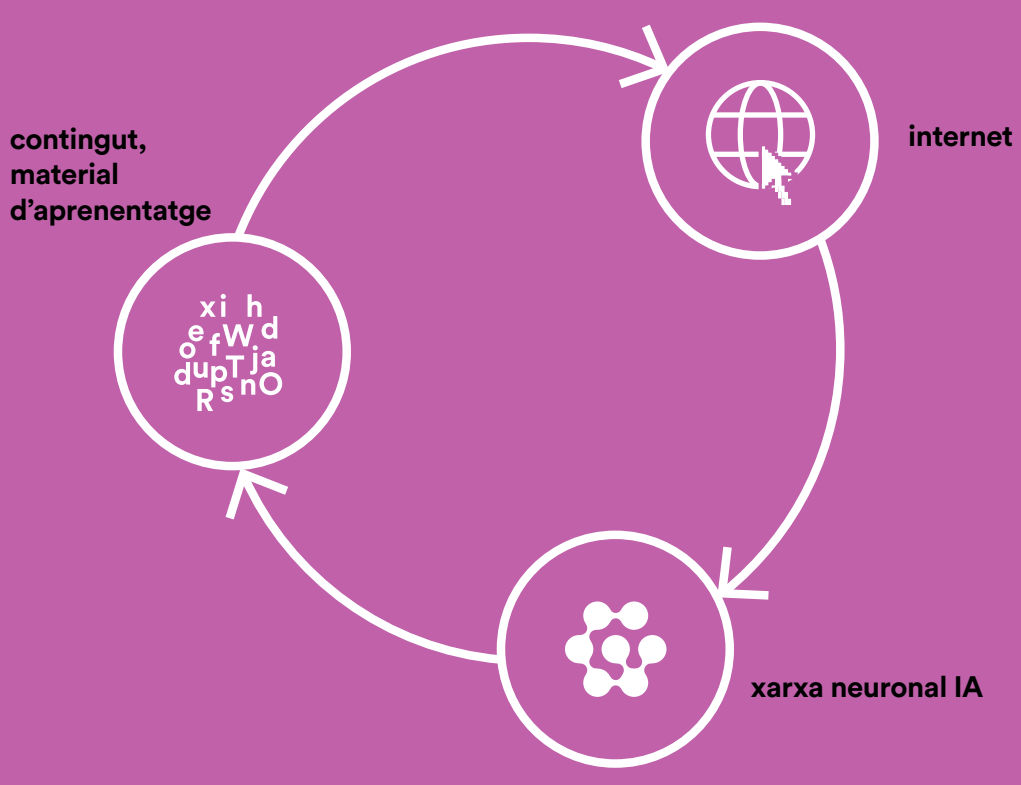
2 Memòria contextual

ChatGPT recorda converses prèvies amb l'usuari. Això fa que les interaccions siguin més naturals, però no sempre millora la precisió, la validesa o la qualitat de les respostes.



3 Col·lapse del model

Si s'entrena el model amb material creat per ell mateix, naturalment la qualitat del sistema s'acaba corrompent.



Quins *conceptes ètics* ens ajuden a analitzar i prevenir aquests riscos?

Antropomorfisme

«Atribució de qualitats i comportaments humans a entitats no humanes; per exemple, quan atorguem emocions o motivacions humanes a animals, màquines, objectes o fenòmens naturals.»

Transparència

«La transparència és el grau en què les dades i els algorismes utilitzats pels sistemes d'IA són accessibles i comprensibles per als usuaris.»

Agència

«Grau d'autonomia, intencionalitat i capacitat de decisió d'un agent, juntament amb el poder i els recursos necessaris per desenvolupar el seu potencial ple.»

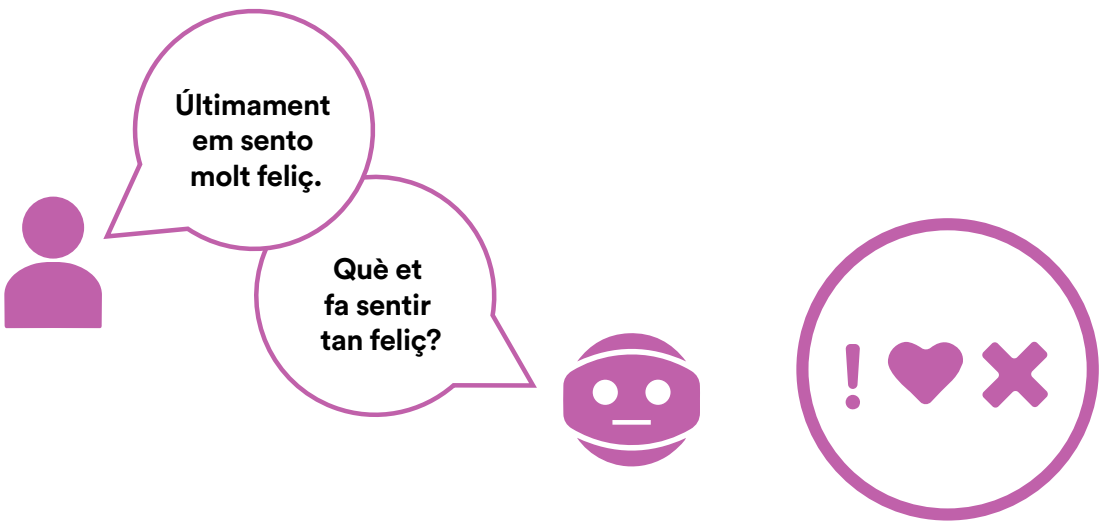
Per què és un problema *l'antropomorfisme*?

Emocions

Atribuir **emocions** pot generar expectatives poc realistes i una confiança i una dependència excessives respecte a la intel·ligència artificial.[5]

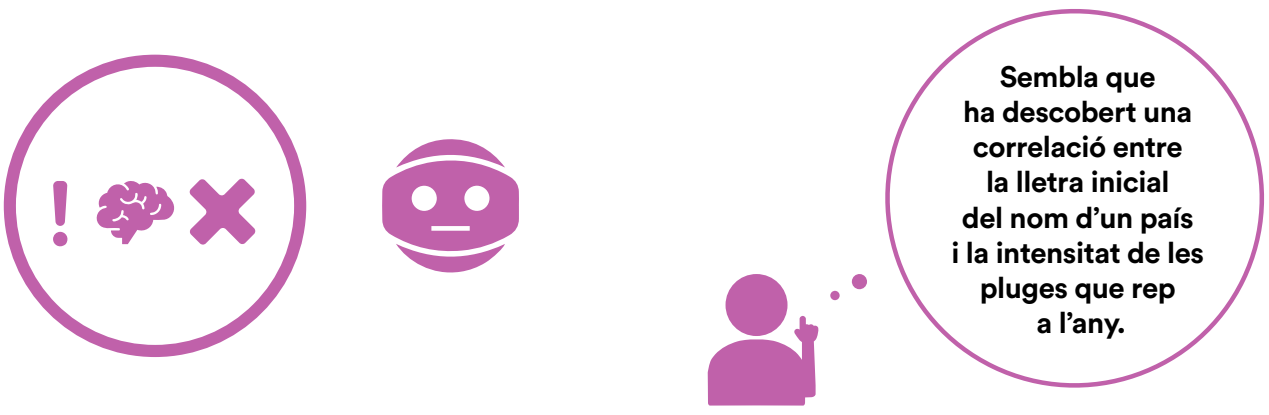
Ja als anys 60, Joseph Weizenbaum, creador d'ELIZA, un senzill *xatbot* psicòleg, va dir del seu software:

«El que no m'esperava és que exposicions extremadament breus a un programa informàtic relativament senzill poguessin induir poderosos pensaments delirants en persones força normals».



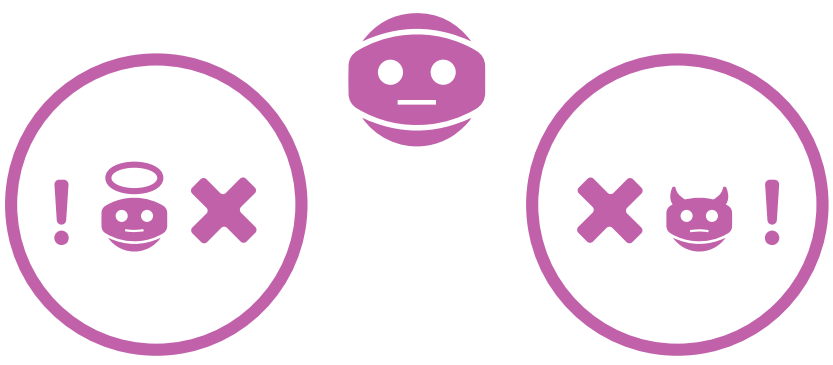
Intel·ligència

Atribuir **intel·ligència** pot tenir conseqüències greus, especialment en la presa de decisions crítiques en què la supervisió humana és crucial.



Responsabilitat

Atribuir **raonament moral** planteja problemes a l'hora de trobar responsabilitats de les decisions.



L'antropomorfisme sembla estar lligat al *joc de la imitació...*

Però... és culpa de Turing que alguns desenvolupadors d'IA creïn màquines que imiten els humans?

No! Turing va proposar que una màquina es pogués considerar intel·ligent quan resultés indistingible d'un humà als ulls d'un avaluador, sempre que aquesta persona estigués preparada i alerta per a la missió.

Però no va suggerir que la veritable intel·ligència s'aconseguís imitant-nos.

De fet, Turing no va afirmar explícitament que el seu test fos una mesura de la «intel·ligència», ja que, a la pràctica, **els resultats depenen més de les actituds, les habilitats o l'enginy de l'avaluador que de la intel·ligència real de la màquina.**[6]

I el cert és que els usuaris no sempre ens mostrem hàbils, enginyosos, preparats i alertes per a la missió d'identificar les característiques pròpies de la intel·ligència humana.

Els usuaris, com totes les persones, tenim una certa tendència innata, generalitzada i ben coneguda, a antropomorfitzar els objectes que ens envolten.

PERÒ NO TOTA LA RESPONSABILITAT ÉS DELS USUARIS...

Els desenvolupadors sovint troben incentius en el fet de donar un toc humà als seus models generatius, i els doten de respostes simpàtiques, educades, sensibles o empàtiques...

Així s'asseguren que l'usuari se sent còmode i escoltat, i passa més temps interaccionant amb l'eina.

És crucial, per tant, que els creadors de tecnologia dissenyin sistemes que **revelin clarament** la seva identitat i, encara més important, que **informin els usuaris** sobre les seves capacitats i limitacions.

Transparència

La transparència és el grau en què les dades i els algorismes utilitzats pels sistemes d'IA són accessibles i comprensibles per als usuaris.

La majoria dels sistemes de comunicació digital que inclouen la intervenció d'una IA generativa no tenen la transparència per als usuaris.

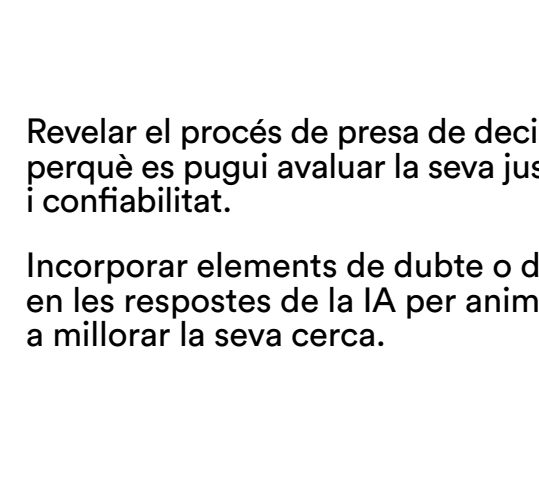
Això significa que, la majoria de vegades que fem servir eines digitals amb IA que crea contingut, no sabem exactament **d'on venen les dades, com funcionen els algorismes** o **en quines condicions treballen les persones** involucrades en el procés de revisió.

Els LLM poden tractar **dades confidencials** sense conèixer la presència d'atributs sensibles, com ara ingressos econòmics o codis postals, cosa que implica un risc de desvelar informació.

La **falta de transparència** en les dades d'entrenament impedeix garantir que aquests sistemes no manegin **informació sensible** en els seus processos, i això posa en dubte la seva **justícia i equitat**.[7]

Anàlisi de transparència

Àmbit	Situació	Accions que augmenten la transparència
DADES	**Origen poc divers**	Mitigar biaixos fomentant, per exemple, l'ús de llenguatges minoritaris en l'entrenament de models.
	Origen no reconegut	Reconèixer la propietat intel·lectual de milions de creadors.
ALGORISMES	**Funcionamient opac**	Revelar el procés de presa de decisions perquè es pugui avaluar la seva justícia, ètica i confiabilitat.
	Resultats dubtosos	Incorporar elements de dubte o d'incertesa en les respostes de la IA per animar l'usuari a millorar la seva cerca.
PERSONES	**El nombre d'etiquetadors s'estima en milions i continua creixent.** [8]	Avaluar la qualitat i la consistència de les dades etiquetades augmenta la fiabilitat i la justícia dels models.
	Entre el 33 % i el 46 % d'ells [9] **utilitzen IA generativa en les seves tasques, amb la qual cosa alimenten el cercle viciós de generació de desinformació i biaix.** [10]	Conèixer i millorar les condicions de treball dels etiquetadors garanteix les pràctiques ètiques. Informar sobre les pràctiques d'inclusió i de diversitat en la selecció de treballadors garanteix la representació equitativa de la població.



Una anàlisi de la transparència dels sistemes d'IA ens fa preguntar-nos:

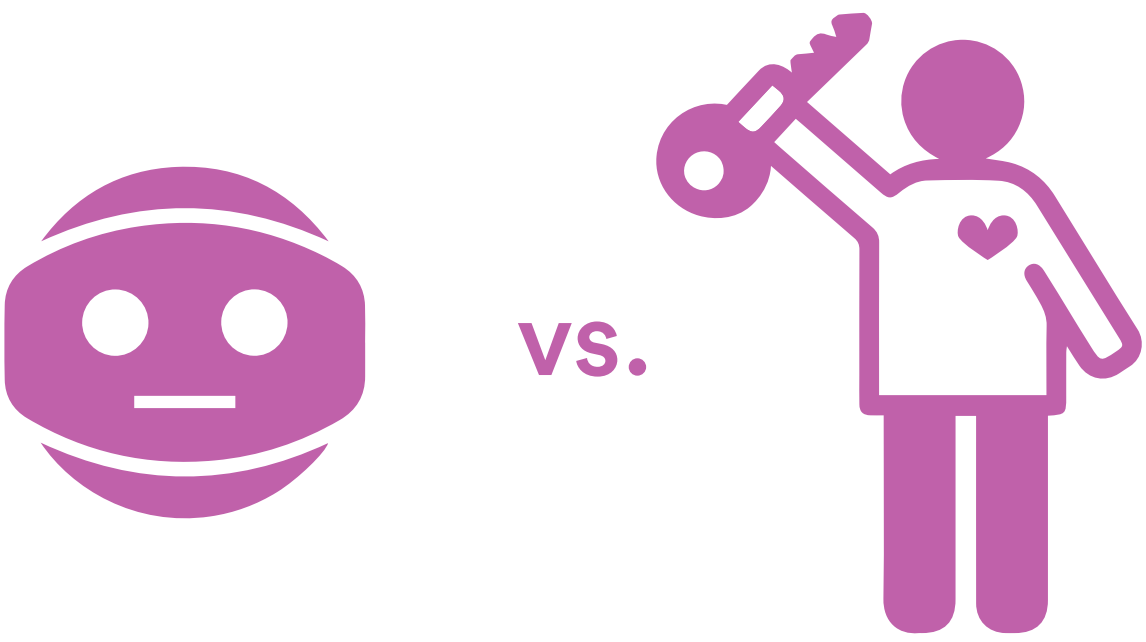
Fins a quin punt la falta de transparència en els sistemes d'IA afecta **la capacitat de les persones per exercir el control** i prendre decisions informades sobre el seu ús?

Agència

L'agència es refereix a la capacitat d'una persona o qualsevol altra identitat per prendre decisions, actuar i tenir control sobre el seu entorn.

PER EXEMPLE, UN FILL GUANYA AGÈNCIA QUAN ELS SEUS PARES LI LLIUREN LES CLAUS DE CASA.

L'agència d'un LLM [11] (què respondre a una pregunta o *prompt*, com escriure o dibuixar, d'on obtenir la informació necessària) és un fenomen nou per a nosaltres: és la primera vegada que veiem que la intel·ligència i l'ús del llenguatge no estan necessàriament vinculats a un cos, una ment i un cor.



En quin sentit és un *nou tipus d'agència*?

Sabem que l'ús del llenguatge es pot descriure amb nombres i imitar amb nombres.[12]

Això significa que les màquines poden generar contingut a partir d'una enorme quantitat de dades.

Però això no significa que les màquines tinguin la nostra agència i autonomia en la facultat d'escriure.

Per què? Què els falta per *ser agents* i *ser autònomes*?

Ni tan *intel·ligent*, ni tan *artificial*[13]

Una primera consideració és que, si les màquines no tenen cos, sentiments, experiència subjectiva…, podem dir que, quan creen, estan expressant alguna cosa?, tenen intenció de comunicar una idea com nosaltres?, entenen el que diuen?

1

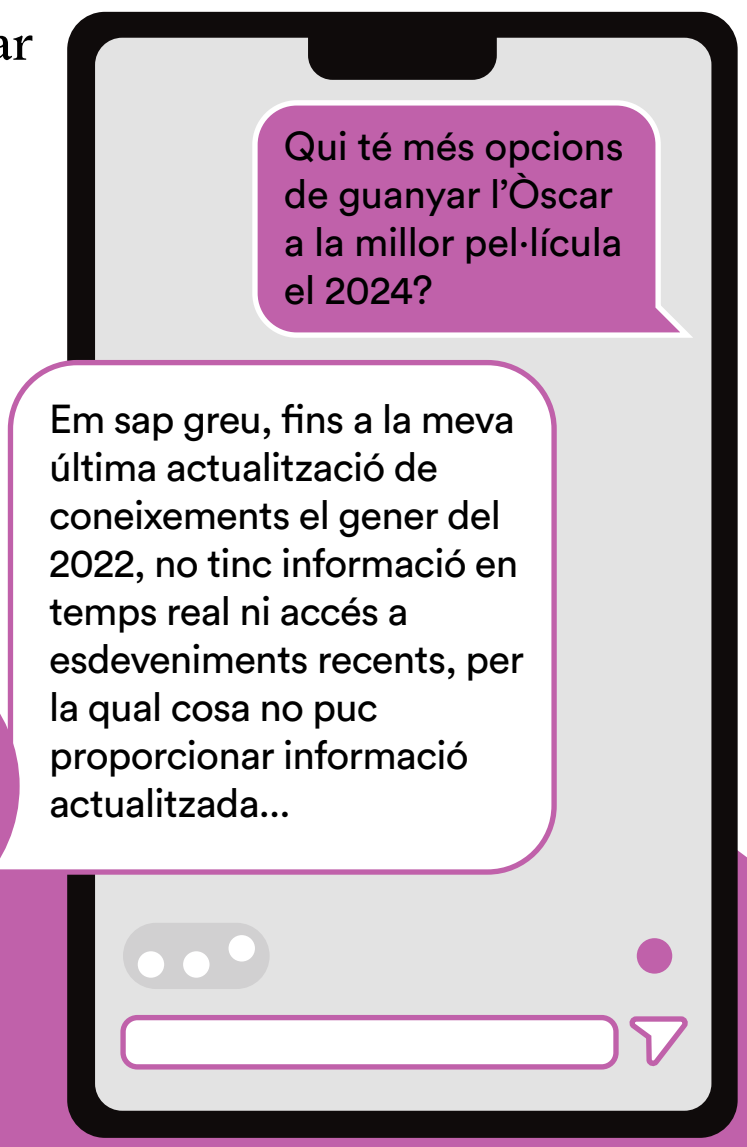
LES MÀQUINES REVELEN LA NOSTRA CAPACITAT CREATIVA, NO LA SEVA.

2

Una altra característica fonamental de la intel·ligència humana és saber reaccionar a situacions noves desconegudes mitjançant l'adquisició constant d'informació.

Però aquesta característica està fora de l'abast dels LLM com ChatGPT perquè, independentment de la quantitat de dades que utilitzen, es troben contínuament amb situacions per a les quals no han estat entrenats.

Per tant, un desafiament per a aquestes tecnologies d'IA és aprendre a interactuar amb les persones per buscar la informació que constantment els falta.



És a dir, *l'agència d'una intel·ligència artificial depèn dels humans*:

Per a la creació de contingut, que depèn de milers de llibres, articles, posts i altre material de milers d'autors (no reconeguts).[14]

Per etiquetar manualment dades —conceptes, imatges, vídeos, categories de text, entitats i intencions darrere de les entrades textuals— fonamentals per a l'entrenament d'algorismes del *xatbot*.

Preguntes de reflexió

1. Quan les màquines escriuen, estan expressant alguna cosa?, tenen la intenció de comunicar una idea com nosaltres?, podem dir que entenen el que volen dir i el que diuen?
2. **Què sents quan estàs davant d'aquest tipus de màquines «intel·ligents»?**
3. Per què els LLM multimodals poden canviar la manera com ens comuniquem?
4. **Consideres l'antropoformització de la tecnologia com un factor de risc? Per què?**
5. Quines són les principals dependències que fan que els LLM no tinguin agència?
6. **Què opines sobre la quantitat de persones que treballen en l'etiquetatge de dades per a la IA i el seu impacte en la indústria?**

REFERÈNCIES PER SABER-NE MÉS

1. Mirzadeh, I. et al., (Apple) (2024). GSM-Symbolic: Understanding the Limitations of Mathematical Reasoning in Large Language Models, https://arxiv.org/pdf/2410.05229
2. Kambhampati, S. (2024). Can large language models reason and plan?. *Annals of the New York Academy of Sciences, 1534*(1), 15-18
3. Hohenstein, J. et al. (2023). Artificial intelligence in communication impacts language and social relationships. *Scientific Reports, 13*(5487)
4. Pennebaker, J. et al. (2003). Psychological aspects of natural language use: Our words, our selves. *Annual review of psychology, 54*, 547-577
5. Stark, L. (2024, juny). Animation and Artificial Intelligence. *Proceedings ofthe 2024 ACM Conference on Fairness, Accountability, and Transparency* (pp. 1663-1671)
6. Proudfoot, D. (2011). Anthropomorphism and AI: Turing's much misunderstood imitation game. *Artificial Intelligence, 175*(5-6), 950-957
7. Anthis, J. et al. (2024). The Impossibility of Fair LLMs. arXiv:2406.03198
8. Birhane, A. (2021). Algorithmic injustice: a relational ethics approach. *Patterns, 2*, 1-9
9. Williams, R. (2023). The people paid to train AI are outsourcing their work to... AI. *MIT Technology Review*, https://www.technologyreview.com/2023/06/22/1075405/the-people-paid-to-train-ai-are-outsourcing-their-work-to-ai
10. Kidd, C., & Birhane, A. (2023). How AI can distort human beliefs. *Science, 380*(6551), 1222-1223
11. Floridi, L. (2023). AI as agency without intelligence: on ChatGPT, large language models, and other generative models. *Philosophy & Technology, 36*(1), 15
12. Shannon, C. E. (1948). A mathematical theory of communication. *The Bell System Technical Journal, 27*(3), 379-423
13. Crawford, K. (2023). *Atlas de la IA. Poder, política y costes planetarios de la inteligencia artificial*. Ned Ediciones
14. Appel, G. et al. (2023). Generative AI Has an Intellectual Property Problem. *Harvard Business Review*, https://hbr.org/2023/04/generative-ai-has-an-intellectual-property-problem

Text: Albert Sabater / Disseny: La Puput Gràfica Coop V
Fundación "la Caixa", 2025