Hauria de tenir consciència una IA? Aproximacions a la consciència humana i a la consciència artificial

Óscar Trabazos Clavería · Assaig, sense revisió per parells · Versió 2.0, 2 de novembre de 2026 (primera versió, 2023)

Resum

Podria una màquina programada per persones tenir experiències subjectives, com veure un color, sentir dolor o gaudir d'una melodia? Aquest assaig aborda aquesta pregunta en quatre parts. La primera repassa què entenem per consciència fenomènica i les respostes clàssiques al problema ment-cos: dualisme, materialisme en les seves diverses variants i idealisme. La segona estudia el cervell com a processador d'informació, des dels inicis de la ciència cognitiva fins als models de memòria i el repte de les metàfores. La tercera ho aplica tot a la intel·ligència artificial: la IA forta, el test de Turing, l'habitació xinesa, les teories científiques actuals de la consciència i els models de llenguatge, amb les seves al·lucinacions i la seva semblança amb les del cervell humà. La quarta afronta les preguntes ètiques i filosòfiques d'una ment artificial, la hipòtesi de l'univers simulat i la relació entre entropia, longevitat i consciència digital.

La conclusió de l'autor és que els sistemes actuals no donen cap prova de consciència: són sistemes d'una capacitat enorme que responen amb la sortida més probable. No podem crear de manera no biològica el que encara no comprenem, i si algun dia una intel·ligència artificial general arriba a alguna cosa semblant, probablement serà una consciència sintètica, diferent de la humana. Les preguntes ètiques, en canvi, ja són actuals i convé fer-se-les abans que la tecnologia les imposi.

Paraules clau: consciència fenomènica, intel·ligència artificial, problema ment-cos, funcionalisme, ciència cognitiva, models de llenguatge, al·lucinacions, ètica de la IA, consciència digital.

Nota sobre les fonts i sobre aquesta edició

Aquest assaig va néixer de les recerques, notes, idees i projectes de l'autor i del seu equip sobre com s'hauria de dissenyar una IA. La primera versió, escrita el febrer del 2023, va prendre com a guia el recorregut teòric de la tesi de màster de Cristóbal Fuentes Barassi, *Conciencia e Inteligencia Artificial: consideraciones críticas sobre la plausibilidad de que una máquina programada posea conciencia fenoménica* (Universidad de Chile, 2011).

Aquesta tesi és la font base de les seccions 2 a 6 i es cita com a tal. Parts d'aquella primera versió es van redactar amb ajuda d'IA. Entre el 2023 i el 2024 l'autor la va divulgar i ampliar en una sèrie del seu blog.

Aquesta versió 2.0 aplega tot aquest treball en un sol text i el posa al dia:

- **Integra la sèrie del blog** (2023-2024): la història de la consciència artificial, els dilemes ètics i filosòfics, les al·lucinacions de la IA i del cervell, l'univers simulat, l'entropia i la consciència digital.
- **Corregeix errors de la primera versió**: dates (La Mettrie és del segle XVIII; el Centre d'Estudis Cognitius de Harvard es va fundar el 1960), l'estructura del model de Baddeley, la distinció entre el problema difícil i l'argument del zombi, què mesura el test de Turing i la diferència entre IA forta i IA estreta. Tres models de memòria que apareixien amb noms inexactes es citen ara amb el seu nom i la seva autoria reals.
- **Hi afegeix el que s'ha investigat des del 2023**: les teories científiques de la consciència i els primers intents d'avaluar-hi els sistemes d'IA actuals.
- **Hi afegeix referències completes** per a cada autor citat.

Les conclusions són de l'autor i surten del que ha escrit sobre el tema; es presenten com a posicions argumentades, no com a resultats demostrats.

Part I. La consciència i la ment

1. Introducció

Amb aquest treball, elaborat a partir de les recerques, notes, idees i projectes que hem fet el meu equip i jo, vull mostrar la complexitat que hi ha darrere de crear una IA i d'adaptar-la a un procés de treball. Adoptar intel·ligència artificial va molt més enllà de connectar-se a una API que resolgui un problema concret. Aquest treball tracta de la necessitat d'entendre com hauria de pensar realment una IA a partir d'entendre com funcionen el nostre cervell i la nostra consciència.

Sense entendre les teories que durant milers d'anys han intentat descobrir el funcionament del nostre cervell, és impossible dir que estem treballant en un projecte d'IA. Ni tan sols és possible pensar en un projecte d'IA que tingui una utilitat real i, sobretot, que sigui ètic.

El treball és llarg. Està pensat per a qui, més enllà de l'aspecte computacional, necessita entendre per què hauria de fer servir tecnologies basades en IA i quin enfocament haurien de tenir els seus dissenys. Avança del més antic al més actual: primer, què és la consciència i com l'ha pensada la filosofia; després, com l'estudia la ciència cognitiva; tot seguit, què significa tot plegat per a la IA d'avui, i finalment, quines preguntes ètiques i filosòfiques ens obre.

Vull subratllar des del principi que amb aquest treball no pretenc semblar espiritual. La meva intenció és recórrer com els éssers humans han intentat entendre la naturalesa de la nostra consciència durant centenars d'anys, perquè és un exercici necessari si ens volem plantejar la possibilitat que una intel·ligència artificial en tingui.

2. La consciència fenomènica

Seria possible crear una màquina programada que tingués consciència fenomènica? Podria un artefacte creat i programat per humans, basat en algorismes, tenir experiències subjectives com veure un color, sentir dolor, percebre una aroma, gaudir d'una melodia o sentir estrès?

2.1 El punt de partida: algorismes i màquines

Des dels treballs fundacionals d'Alan Turing el 1948 i el 1950, la intel·ligència artificial ha intentat crear màquines capaces d'emular les capacitats cognitives humanes i de competir-hi, amb el concepte d'algorisme com a pedra angular. Un algorisme és una sèrie ordenada i finita de passos que porta a un resultat, i és el que hi ha darrere de l'anomenada màquina de Turing: una màquina teòrica que té diferents estats i que, donada una entrada, passa a l'estat següent. La idea d'aquesta aproximació clàssica és que, si es dota una màquina de l'algorisme adequat, hauria de poder fer qualsevol tasca que pugui fer un ésser humà.

D'aquest optimisme va néixer l'interès per crear màquines que imitessin altres característiques del pensament humà, i en particular la consciència. És el que es coneix com a consciència artificial, també anomenada consciència de màquina o consciència sintètica. Els seus motius es poden agrupar en quatre:

1. Dissenyar models de consciència per identificar-ne els trets determinants, com a part de la feina de la ciència cognitiva.
2. Fer servir les implementacions d'aquests models com a eines per posar a prova i entendre la consciència humana.
3. Determinar com és la relació entre la consciència i el suport físic en què es produeix.
4. Comprendre que la consciència fenomènica és fonamental en algunes de les tasques que considerem intel·ligents.

La literatura i el cinema han explorat durant dècades la idea de màquines capaces de tasques complexes: HAL 9000 a *2001: una odissea de l'espai*, els robots de *Jo, robot* d'Isaac Asimov o les màquines conscients d'*A.I. Intel·ligència Artificial* de Steven Spielberg. A la vida real, assistents de veu com Siri o Alexa aprenen dels seus usuaris i responen a les seves ordres, sense que ningú els atribueixi per això una vida interior.

Tot i que la disciplina ha creat màquines amb capacitats cognitives similars o superiors a les humanes en tasques concretes, continua oberta la pregunta de si poden tenir experiències subjectives, com assaborir un te o gaudir d'una música. I amb ella, d'altres: com saber si una màquina en té, quina relació hi ha entre el suport físic i la seva aparició, i si les màquines pensen o només calculen.

Deep Blue, l'ordinador que va vencer Kasparov el 1997, n'és un bon exemple. Juga als escacs millor que gairebé qualsevol humà, però res en el seu disseny no li permet viure l'experiència de jugar: ni la tensió de la partida, ni la pressió de la competició, ni la inquietud davant del moviment del rival. Segueix un procediment; no li passa res.

2.2 Per què la conducta no basta

Els defensors de l'anomenada IA forta sostenen que n'hi ha prou d'observar el comportament extern d'una màquina per atribuir-li intel·ligència. És el que se sol associar al test de Turing, tot i que Turing el va proposar per a la intel·ligència, no per a la consciència (secció 6.2). I és, en el fons, el que fem cada dia amb altres persones: suposem que són conscients perquè es comporten com nosaltres i se'ns assemblen.

El problema és que la consciència fenomènica només l'experimenta, en primera persona, qui la té. Una prova objectiva de conducta ens pot dir què fa una màquina, no si sent alguna cosa en fer-ho. Per això l'afirmació conductista que la consciència es pot detectar amb proves externes no resol la qüestió. El problema no és exclusiu de la intel·ligència artificial: és el vell problema ment-cos de la filosofia, que ha donat lloc al dualisme, al monisme i a moltes posicions intermèdies, i sobre el qual encara no hi ha consens.

2.3 Què entenem per consciència

Bernard Baars, psicòleg i neurocientífic cognitiu i autor de la teoria de l'espai de treball global (secció 6.4), assenyala la confusió que produeix plantejar la consciència en termes de subjectivitat, de «què se sent en ser algú»: s'acaba en una de les tres posicions clàssiques, el mentalisme, el fisicalisme o el dualisme. Per avançar, proposa formular les preguntes sobre la consciència de manera que es puguin respondre, amb un llenguatge i un enfocament concrets.

La formulació més citada d'aquesta subjectivitat és de Thomas Nagel (1974): un organisme té estats conscients si hi ha alguna cosa que se sent en ser aquest organisme. El seu exemple, «què se sent en ser un ratpenat?», mostra que podem saber-ho tot sobre l'ecolocalització d'un ratpenat sense saber com és, per a ell, percebre així el món.

Ned Block (1995) adverteix que «consciència» és un concepte híbrid: una sola paraula per a fenòmens diferents, com l'atenció, adonar-se d'alguna cosa externa o interna, la identitat personal o l'experiència de colors i dolors. En distingeix sobretot dos: la

consciència fenomènica, que és l'experiència, i la consciència d'accés, que és la informació disponible per raonar, parlar i guiar la conducta. En les persones solen anar juntes, però són conceptes diferents.

David Chalmers (1995) va posar nom a la dificultat central: el problema difícil de la consciència. Els «problemes fàcils», com discriminar estímuls, integrar informació o informar dels propis estats, es poden explicar descrivint funcions i mecanismes. El difícil és explicar per què aquests processos van acompanyats d'experiència. Per il·lustrar-ho, Chalmers (1996) fa servir un argument diferent, el del zombi filosòfic: un ésser físicament idèntic a una persona, que es comporta exactament igual, però que no sent res. Si aquest ésser és tan sols concebible, la consciència no es redueix al que és físic ni a la conducta. Chalmers ha arribat a considerar seriosament el panpsiquisme, la idea que la consciència és una propietat fonamental de la natura.

2.4 L'abast d'aquest assaig

En aquest text analitzo críticament la possibilitat que una màquina programada amb algorismes tingui consciència fenomènica, és a dir, l'experiència subjectiva del món que viu un ésser en relacionar-se amb el seu entorn i amb la seva vida interior. Hi ha altres nocions de consciència, com l'autoconsciència o la consciència d'accés, però l'anàlisi se centra en aquesta, perquè és la que fa difícil la pregunta.

Per fer-ho cal discutir els fonaments filosòfics sobre els quals treballen disciplines com la IA i la psicologia, i tenir presents els interrogants ètics i legals que obre. La tesi que serveix de base a aquest recorregut (Fuentes Barassi, 2011) sosté que dotar una màquina de consciència fenomènica seria possible i necessari, perquè aquesta capacitat és fonamental per a certes tasques intel·ligents i no un simple epifenòmen; però dubta que el modelatge mitjançant algorismes sigui el camí, perquè els algorismes no expliquen com ni per què es produeix l'experiència. La meva posició, que desenvolupo a les conclusions, és més prudent amb el primer punt i coincideix amb el segon.

Una alternativa a l'algorisme clàssic que s'ha proposat és construir xarxes neuronals artificials que simulin les xarxes neuronals biològiques. La seva dificultat és doble: no sabem quins trets de la xarxa biològica són els que importen per a la consciència, i no tenim manera d'avaluar si una xarxa artificial en té.

2.5 Primeres conclusions sobre la consciència fenomènica

La consciència fenomènica és l'experiència de «què se sent»: el gust del menjar, un color brillant, la sensació d'estar en el món i de tenir una experiència pròpia dels objectes, els fets i els pensaments. No es pot mesurar des de fora: és única per a cada individu, i és el que ens fa conscients de nosaltres mateixos i del que ens envolta. Per això se la considera l'aspecte més misteriós de la ment.

Es distingeix de la consciència d'accés, que és la capacitat de disposar de la informació per fer-la servir en decisions, i de l'autoconsciència, que és la capacitat de saber-se un ésser separat i individual.

Sobre la IA hi ha dues postures de fons. Uns creuen que és possible programar una màquina perquè tingui consciència fenomènica, i que per fer-ho la màquina hauria de poder sentir i experimentar el món de manera semblant a com ho fem nosaltres. D'altres sostenen que la consciència és una propietat emergent de la complexitat de la matèria i de l'organització biològica del cervell, i que requereix un substrat biològic que cap codi no pot substituir. No hi ha consens, i entendre per què exigeix mirar el problema ment-cos.

3. El problema ment-cos

El problema ment-cos és un dels temes centrals de la filosofia de la ment: com es relacionen la ment (pensaments, sentiments, consciència) i el cos (cervell, sistema nerviós). Sorgeix perquè sembla que hi ha una diferència de fons entre tots dos. La ment és conscient i té experiències subjectives; el cos és un objecte físic que es pot estudiar amb les ciències naturals. Com interactuen dues coses tan diferents?

3.1 Quatre exemples

Per explicar-me millor, deixeu-me fer servir alguns exemples.

- **Una pel·lícula de terror.** La imatge i el so arriben als ulls i a les orelles i el cervell els processa. Però a més tens una experiència subjectiva de por. Com dóna lloc l'activitat elèctrica i química de les teves neurones a aquesta por que sents?
- **La mà a l'estufa.** Una persona es crema la mà. El contacte amb l'estufa i el dany als teixits són fets físics; el dolor que sent només el pot experimentar ella. Com es tradueix l'activitat de les neurones de la mà i del cervell en aquesta experiència de dolor?
- **Una poma vermella.** Els ulls i el cervell processen la informació visual i la persona reconeix que veu una poma vermella. Però a més hi ha alguna cosa que se sent en veure aquest vermell.
- **Una música.** Les orelles i el cervell processen el so i la persona reconeix la cançó. Però a més en gaudeix, o la pateix.

En els quatre casos, la ciència pot descriure el processament amb un detall cada vegada més gran. El que no aconsegueix explicar és el pas del processament a l'experiència. La ment és una cosa separada del cos, o una propietat que emergeix del cervell? Les grans respostes són tres: el dualisme, el materialisme i l'idealisme.

3.2 Dualisme

El dualisme defensa que a l'univers hi ha dos tipus de substància: la matèria i la ment o l'esperit. La ment no seria material ni es podria reduir a processos físics, químics o biològics del cervell; seria una entitat independent, que podria existir sense el cos, encara que hi interactuï.

La seva formulació clàssica és la de René Descartes al segle XVII: la ment és una *res cogitans*, una substància pensant, i el cos és una *res extensa*, una substància extensa que es pot mesurar i estudiar. Per al dolor de la mà cremada, el dualista dirà que no s'explica només en termes físics: requereix una explicació mental addicional.

El dualisme respon al problema ment-cos, però no el resol: la relació entre les dues substàncies continua sent un misteri. Ha rebut tres crítiques principals. No explica com una cosa immaterial pot moure una cosa material. Ofereix una visió fragmentada de l'ésser humà. I sembla xocar amb la física, que no deixa espai per a causes no físiques.

Per a la IA, el dualisme planteja un repte de fons: si la ment no és només un producte del cervell, com es pot replicar en una màquina? Molts dualistes sostenen que la IA no arribarà mai a una veritable consciència, perquè aquesta va més enllà de la funció i implica una dimensió subjectiva: la màquina podrà simular la intel·ligència, però no saber què és ser conscient. D'altres creuen que el dualisme és massa limitat per a aquest debat i proposen alternatives com el funcionalisme o el connexionisme, que expliquen la ment per processos cognitius i neuronals.

3.3 Materialisme

El materialisme sosté el contrari: tot el que existeix és matèria i energia, i la ment i la consciència són el resultat de processos materials i biològics del cervell. La consciència no és una entitat a part, sinó una propietat o una funció del cervell que emergeix del seu funcionament. En la seva versió més radical, el monisme materialista, la ment és un epifenòmen de l'activitat neuronal, i no existeix un «dolor» mental independent d'aquesta activitat.

Per a la IA, la conseqüència és directa: si entenem prou bé el cervell, podríem construir una màquina que en reproduís els processos i les propietats que en sorgeixen, sense postular res immaterial. I la consciència es podria estudiar amb la ciència empírica, en particular la neurociència i la psicologia cognitiva.

El materialisme filosòfic es remunta a l'antiga Grècia, on Demòcrit i Epicur van sostenir que tot està fet d'àtoms i que la ment i el cos són inseparables. Al segle XVIII, el metge i filòsof francès Julien Offray de La Mettrie va defensar a *L'home màquina* (1747) que el cos humà és una màquina complexa i la ment una funció del cervell. Després el van defensar, amb matisos molt diferents, Ludwig Feuerbach, Karl Marx i Friedrich Engels, i en la filosofia de la ment actual Daniel Dennett i Patricia Churchland.

Cal recordar que la filosofia de la ment no és una ciència empírica: reflexiona sobre els conceptes i els supòsits amb què la ciència estudia la ment. Però aquestes posicions condicionen com s'investiga, i la neurociència i la ciència cognitiva han generat al seu torn els seus propis debats. Del tronc materialista en surten diversos corrents que convé distingir.

3.3.1 Conductisme

El conductisme es va desenvolupar als Estats Units a la primera meitat del segle XX. Estudia només la conducta observable i mesurable i evita qualsevol referència a processos mentals interns. El comportament s'explica per estímuls i respostes, i les recompenses i els càstigs són els mecanismes que el modelen.

El seu fundador, John B. Watson, volia una psicologia objectiva i experimental, sense introspecció: l'ésser humà com una màquina el comportament de la qual resulta de l'entorn i de les respostes apreses. B. F. Skinner va desenvolupar el condicionament operant, segons el qual qualsevol conducta, fins i tot la més complexa, s'explica per reforços i càstigs.

El conductisme va dominar la psicologia durant dècades, però avui es considera massa simple per explicar la complexitat humana. Per a la nostra pregunta, deixa fora just el que es pregunta: una màquina que es comporta com si sentís seria, a efectes conductistes, igual que una que sent.

3.3.2 Funcionalisme

El funcionalisme entén els estats mentals per la seva funció: pel que hi entra, el que en surt i el processament que hi ha entre una cosa i l'altra. Un mal de cap seria un estat causat per certs estímuls que produeix certes conductes, com prendre un analgèsic o evitar la llum. El que defineix l'estat no és el material de què està fet, sinó el paper que fa. Hilary Putnam (1967) el va formular en filosofia de la ment amb la idea que un mateix estat mental es pot realitzar en suports físics diferents.

Davant del conductisme, el funcionalisme reconeix els processos mentals interns com a necessaris per explicar la conducta. L'han adoptat la psicologia cognitiva, la neurociència cognitiva i la intel·ligència artificial, que busca sistemes capaços de percebre, raonar i decidir sense replicar la complexitat biològica del cervell. És la posició que fa pensable una màquina conscient, i per això hi tornarem quan parlem de la IA forta.

3.3.3 Fisicalisme

El fisicalisme sosté que tot el que existeix és físic o es pot explicar en termes físics. En filosofia de la ment, vol dir que la ment és un producte del cervell i de l'activitat neuronal, i no una entitat independent del cos. Els seus defensors es recolzen en la neurociència, que

explica cada vegada més processos mentals que abans semblaven fora de l'abast de la física.

Se li fan dues crítiques principals. La primera és la mateixa consciència: l'experiència subjectiva sembla resistir-se a l'explicació física, com mostra el problema difícil. La segona és la intencionalitat: el fet que els pensaments siguin *sobre* alguna cosa i tinguin significat.

3.3.4 La teoria de la identitat

Un dels enfocaments més destacats del fisicalisme és la teoria de la identitat, formulada per U. T. Place (1956) i J. J. C. Smart (1959): els estats mentals són idèntics a estats del cervell. Quan algú sent felicitat, el que passa és una activitat neuronal concreta; quan sent dolor, una altra.

El seu exemple clàssic és «el dolor és l'estimulació de les fibres C». Les fibres C són fibres nervioses sense mielina que transmeten lentament el dolor persistent i la temperatura; les fibres A, mielinitzades, transmeten més ràpid el dolor agut i la pressió. Els filòsofs van fer servir aquest exemple com a simplificació: avui sabem que el dolor implica xarxes cerebrals àmplies, però la idea de fons és la mateixa. El dolor no seria una cosa diferent de l'activitat física, sinó una forma particular d'aquesta activitat. Portada a la seva conclusió, «la ment és el cervell»: no hi ha cap entitat anomenada ment a part de l'activitat cerebral.

La teoria de la identitat va ser molt influent a la segona meitat del segle XX. Després va perdre terreny davant del funcionalisme, precisament per l'argument que un mateix estat mental es podria donar en cervells o suports diferents. Per a la IA la diferència és crucial: si la consciència és idèntica a un tipus d'activitat biològica, una màquina de silici no la podria tenir; si és una funció, potser sí.

3.4 Idealisme

L'idealisme sosté que la realitat material no existeix de manera independent de la ment que la percep: el que podem conèixer és una construcció mental. Hi ha dos grans corrents. L'idealisme objectiu atribueix la realitat a un principi racional o diví; el subjectiu, a l'activitat cognitiva i perceptiva del subjecte.

Els seus exponents clàssics són Plató, per a qui el que percebem és ombra o còpia d'un món d'idees eternes; George Berkeley, per a qui les coses només existeixen en la mesura que són percebudes; i, amb plantejaments molt diferents, Kant i Hegel.

Avui l'idealisme té poc predicament entre científics i filòsofs de la ment, perquè no ofereix criteris que es puguin posar a prova. Alguns autors han intentat adaptar-lo a la mecànica quàntica, amb l'argument que la realitat quàntica només es concreta en ser observada o mesurada, però aquesta lectura és minoritària entre els físics. Per a la IA aporta poc: no diu què cal construir ni com comprovar-ho.

3.5 Síntesi: què implica cada posició per a una màquina

Posició	Què sosté	Què implica per a una màquina
Dualisme	La ment és una substància diferent del cos	Programar un suport físic no bastaria per crear una ment
Conductisme	Només compta la conducta observable	Una màquina que actua com si sentís seria igual, però es perd l'experiència
Teoria de la identitat	Cada estat mental és un estat del cervell	La consciència estaria lligada al substrat biològic; una màquina de silici no la tindria
Funcionalisme	Un estat mental es defineix per la seva funció	Si es reprodueix la funció, el material no importa: és la base de la IA forta
Fisicalisme en general	Tot és físic o explicable físicament	Una màquina conscient seria possible en principi, si sabéssim què cal reproduir
Idealisme	El mental és primer	No dóna criteris per construir ni per comprovar

Part II. El cervell com a processador

4. La ciència cognitiva

La ciència cognitiva és una disciplina interdisciplinària que estudia els processos mentals: la percepció, l'atenció, la memòria, el llenguatge, el pensament i la resolució de problemes. Li interessa el problema ment-cos perquè vol entendre com funciona la ment, com es relaciona amb el cervell i el cos, i com aquesta relació produeix l'experiència conscient.

Treballa amb enfocaments molt diferents. La psicologia cognitiva estudia els processos mentals mitjançant l'observació i l'experiment amb persones. La neurociència cognitiva investiga com sorgeixen de les estructures i els processos del cervell. I la intel·ligència artificial intenta reproduir-los amb algorismes i models computacionals. Per sobre de tots hi ha el computacionalisme, que considera la ment un sistema de processament d'informació.

4.1 Els inicis

La ciència cognitiva va néixer entre els anys quaranta i cinquanta com a resposta a la crisi del conductisme, que havia deixat fora els processos mentals interns. El naixement de la intel·ligència artificial en aquells mateixos anys la va impulsar: per construir sistemes intel·ligents calia entendre quins processos mentals fan falta.

Els primers treballs es van centrar en la percepció visual i auditiva i en el processament del llenguatge, combinant experiments d'estímul i resposta amb registres fisiològics com l'electroencefalografia (EEG). El 1960, els psicòlegs George Miller i Jerome Bruner van fundar el Centre d'Estudis Cognitius de Harvard, que es va convertir en un dels seus focus principals. John von Neumann, matemàtic i pioner de la computació, va comparar el cervell amb l'ordinador a *The Computer and the Brain* (1958), i aquesta analogia va consolidar l'enfocament computacional. En les dècades següents el camp es va estendre a la memòria, el raonament i la presa de decisions, i tècniques com la ressonància magnètica funcional (fMRI) i la tomografia per emissió de positrons (PET) van permetre observar l'activitat cerebral associada a cada procés.

Els pioners de la IA, com Turing i von Neumann, van proposar que el cervell funciona com una màquina de processar informació i que la ment es pot descriure com un conjunt de processos computacionals. Aquesta idea va permetre modelar processos mentals complexos i fer prediccions sobre la conducta en tasques molt variades, i va portar la ciència cognitiva a la psicologia, la neurociència, la filosofia i la lingüística.

En definitiva, la ciència cognitiva intenta entendre com funciona realment el nostre cervell i com processa i organitza la informació. És, per tant, la disciplina més propera a com s'hauria de pensar i organitzar una IA. Sense ella no entendríem com dissenyar una IA ni quina hauria de ser la seva finalitat real.

5. El cervell: processador i sistema de dades biològic

5.1 Computacionalisme

El computacionalisme sosté que el cervell és una mena d'ordinador que processa informació mitjançant algorismes i estructures de dades. Els estímuls sensorials es transformen en representacions mentals, que s'emmagatzemen a la memòria i guien la conducta; aquestes representacions es combinen i es manipulen amb algorismes, i així el cervell calcula, resol problemes i decideix.

Ha estat una perspectiva molt fèrtil: ha donat models computacionals de la percepció visual, de la memòria o de la presa de decisions. Però també ha rebut crítiques. Molts sostenen que el cervell és massa complex per modelar-lo com un ordinador i que la ment

és molt més que un processador. I la idea que la informació és la matèria primera del cervell xoca amb aspectes de la cognició que depenen més de l'experiència, el cos i la cultura que de la informació en estat pur.

5.2 Models cognitius de la memòria

Els models cognitius, o arquitectures cognitives, descriuen com el cervell processa i organitza la informació: com l'adquirim, la desem i la fem servir. Es componen de mòduls que treballen junts, i també s'han fet servir per construir sistemes que imiten el processament humà, en la IA i en la robòtica cognitiva. La memòria n'és el millor exemple.

El model d'Atkinson i Shiffrin (1968)

Descriu la memòria com un flux d'informació per tres magatzems:

1. **Memòria sensorial.** Reté la informació dels sentits durant una fracció de segon. Té una part icònica, visual, i una d'ecoica, auditiva.
2. **Memòria a curt termini.** Manté la informació uns 15 a 30 segons, amb una capacitat limitada. És on s'hi treballa activament.
3. **Memòria a llarg termini.** El magatzem més gran i durador, de capacitat pràcticament il·limitada.

Va ser una contribució decisiva, perquè va donar una descripció detallada del recorregut de la informació. Després es va criticar per massa lineal, i es van proposar models més dinàmics que tenen en compte l'atenció, l'emoció i la motivació.

El model de Baddeley i Hitch (1974)

Alan Baddeley i Graham Hitch van substituir el magatzem únic a curt termini per una memòria de treball amb diversos components:

- **El bucle fonològic**, per a la informació verbal i auditiva. Té un magatzem fonològic passiu, que reté sons uns quants segons, i un procés de repàs articulatori que els manté repetint-los mentalment.
- **L'agenda visuoespacial**, per a la informació visual i espacial.
- **L'executiu central**, que supervisa i coordina els altres components i controla l'atenció.

Baddeley (2000) hi va afegir un quart component, el búffer episòdic, que integra la informació dels altres sistemes i de la memòria a llarg termini en episodis coherents. El model ha estat àmpliament acceptat i també es fa servir per estudiar trastorns de la memòria.

Altres models

A partir d'aquí van sorgir models més complexos, que incorporen l'atenció, l'emoció i la motivació. En destaco quatre línies:

- **La comparació de models de memòria de treball de Miyake i Shah (1999).** Més que un model propi, el seu llibre *Models of Working Memory* va confrontar les principals teories i va mostrar en què coincideixen: un emmagatzematge limitat per al material verbal i visuoespacial i un control executiu que el gestiona.
- **La memòria emocional de LeDoux (1996).** La informació amb càrrega emocional segueix una via ràpida i automàtica, del tàlem a l'amígdala, abans que l'escànça la processi amb més detall per una via més lenta.
- **La integració de cognició i emoció de Pessoa (2013).** Sosté que l'emoció i la cognició no són sistemes separats, sinó que s'integren al cervell, i que el control executiu permet modular l'atenció i l'emoció segons la tasca i el context.
- **La memòria autobiogràfica de Conway i Pleydell-Pearce (2000).** Els records autobiogràfics són construccions dinàmiques que integren informació episòdica, semàntica i emocional, organitzades al voltant dels objectius del jo; recuperar-los depèn de la seva accessibilitat, de la seva coherència i de la capacitat de situar-los en el temps i l'espai.

5.3 En quin punt som

Els models més recents són més dinàmics i tenen en compte més factors, gràcies a tècniques que permeten observar millor el cervell. Destaquen tres idees, cadascuna amb la seva teoria de referència:

- **La memòria de treball es pot ampliar amb la pràctica.** Ericsson i Kintsch (1995) van mostrar que els experts en un camp fan servir estructures de la memòria a llarg termini com si fossin memòria de treball, cosa que explica la seva enorme capacitat en el seu domini. L'atenció selectiva i la inhibició són claus per a aquest funcionament.
- **La memòria és una xarxa de relacions.** La teoria de la memòria relacional (Cohen i Eichenbaum, 1993) sosté que l'hipocamp desa relacions entre elements, i que aquestes relacions són essencials per recuperar i fer servir la informació amb flexibilitat.
- **La memòria va evolucionar per adaptar-nos.** La recerca sobre memòria adaptativa (Nairne et al., 2007) mostra que recordem millor el que és rellevant per a la supervivència: la memòria és selectiva i està esbiaixada cap al que importa per al benestar.

A aquestes tres cal afegir-hi una idea que en l'última dècada ha guanyat molt pes: la del cervell predictiu (Clark, 2013). Segons aquesta idea, el cervell no es limita a rebre informació, sinó que prediu contínuament el que percebrà i corregeix les seves

prediccions amb el que arriba dels sentits. És una idea important per a aquest assaig, perquè els models de llenguatge actuals també són, a la seva manera, màquines de predir (secció 7).

5.4 El gran repte: les metàfores

Però sens dubte un dels grans reptes per entendre la nostra ment és a les metàfores que som capaços de crear, comunicar i entendre. No hi ha una teoria única que les expliqui, però diverses teories de la cognició i la lingüística ofereixen enfocaments complementaris:

- **Lakoff i Johnson (1980)** sostenen que les metàfores no són un ornament del llenguatge, sinó la manera com conceptualitzem el món: entenem el que és abstracte per analogia amb el que és concret i corporal («ser al cim», «una idea que cau»).
- **La teoria de la rellevància de Sperber i Wilson (1986)** explica com inferim el que l'altre vol dir buscant la interpretació més rellevant amb el mínim esforç.
- **La integració conceptual de Fauconnier i Turner (2002)** descriu com combinem espais mentals diferents per crear significats nous.
- **La simulació corporitzada de Gibbs (2006)** proposa que entendre una metàfora implica simular, en part, l'experiència corporal a què remet.

Les metàfores mostren un límit del computacionalisme: bona part del nostre significat està ancorat en el cos i en l'experiència viscuda. És una de les raons per les quals entendre com processem la informació no basta per entendre com sentim. Tots aquests models descriuen la consciència d'accés, el que fem amb la informació; cap no explica encara per què recordar una cosa, o entendre una metàfora, se sent d'una manera determinada.

6. La intel·ligència artificial i la consciència

Després de revisar les teories que intenten explicar per què tenim consciència i com funciona la nostra ment, vegem com les assumeix la IA i com es planteja l'ambiciós objectiu d'una consciència pròpia. Si és que és possible, perquè fins avui no ha passat.

6.1 IA forta, IA feble i intel·ligència general

Convé separar tres idees que sovint es barregen:

- **IA feble o estreta:** sistemes que resolen tasques concretes, de vegades millor que qualsevol persona, sense pretendre comprendre ni sentir.
- **Intel·ligència artificial general (AGI):** un sistema capaç d'igualar o superar les persones en pràcticament qualsevol tasca cognitiva, d'aprendre, raonar i adaptar-se a situacions noves de manera autònoma, com ho fa un ésser humà. Es refereix a l'amplitud del que un sistema pot fer.

- **IA forta**, en el sentit en què la va definir John Searle (1980): la tesi que un ordinador programat adequadament no simula una ment, sinó que *té* una ment, amb estats cognitius reals. Es refereix a si el sistema pensa i sent de debò.

La IA forta es recolza en el funcionalisme: el que importa de la ment no és el substrat físic, sinó la funció que fa, i per tant una ment es podria implementar en un suport diferent del cervell sempre que fes les mateixes funcions cognitives i conductuals. D'aquí que alguns filòsofs i científics hagin explorat la possibilitat d'una IA amb una ment comparable a la humana.

Avui la majoria dels experts consideren l'AGI un objectiu encara no assolit, i la IA forta, una qüestió oberta. Els sistemes actuals superen les persones en tasques concretes, com els escacs o l'anàlisi de grans quantitats de dades, però tenen límits importants en la comprensió profunda del llenguatge, el raonament abstracte, la creativitat i, sobretot, la consciència.

Els projectes que se solen citar com a exemples d'IA avançada són, en realitat, IA estreta: els sistemes de diagnòstic mèdic que analitzen grans volums de dades clíniques; els vehicles autònoms que perceben el trànsit i decideixen en temps real; els programes que juguen als escacs o al go i guanyen, i els sistemes de detecció de frau que analitzen patrons de comportament en transaccions financeres. Resolen molt bé una tasca, i cap no pretén tenir experiència.

6.2 El test de Turing i l'habitació xinesa

Alan Turing, considerat el pare de la informàtica moderna, va ser el primer a plantejar seriosament si les màquines poden pensar. El 1950 va proposar substituir aquesta pregunta, que considerava massa vaga, per un joc d'imitació: si una màquina conversa de manera que un observador no la pot distingir d'una persona, la podem considerar intel·ligent. El test va desafiar la noció tradicional d'intel·ligència, però no abordava la consciència. És un criteri de conducta, i deixa intacte el problema difícil.

La crítica més coneguda a la IA forta és l'argument de l'habitació xinesa de Searle (1980). Imaginem una persona que no sap xinès, tancada en una habitació amb un manual de regles en la seva llengua. Per una escletxa li arriben preguntes escrites en xinès; seguint el manual, compon respostes correctes i les torna. Des de fora sembla que a l'habitació hi ha algú que entén el xinès, però la persona no n'entén ni una paraula: només manipula símbols segons regles. Searle conclou que això és el que fa un programa, i que la sintaxi no basta per a la semàntica: executar el programa adequat no dóna comprensió, i encara menys experiència.

6.3 Una història d'optimisme i escepticisme

La consciència en la IA ha fascinat i dividit des dels primers dies de la computació. Als anys seixanta i setanta hi va haver un gran optimisme: molts creien que una IA amb capacitats humanes era a tocar. Aquest optimisme va xocar amb dificultats tècniques i teòriques, van arribar els anomenats hiverns de la IA, i arguments com el de l'habitació xinesa en van fer visibles els límits: no només es dubtava que les màquines poguessin pensar, sinó que poguessin sentir.

El segle XXI ha portat un renaixement impulsat per nous algorismes i maquinari. Les xarxes neuronals i l'aprenentatge profund han permès a les màquines reconèixer imatges o processar el llenguatge, tasques que abans es consideraven exclusivament humanes, i això ha reavivat el debat. Pot una màquina que aprèn i s'adapta ser també conscient? Hi ha dues grans postures. Per a uns, la consciència és un producte de la complexitat, i sorgirà a mesura que les màquines siguin més complexes. Per a d'altres, és pròpia dels éssers vius, i cap màquina no la podrà tenir, per avançada que sigui.

6.4 Les teories científiques actuals de la consciència

Des que vaig escriure la primera versió d'aquest treball, el debat ha passat de l'especulació a un intent d'examen sistemàtic. La neurociència ha desenvolupat diverses teories de la consciència que proposen mecanismes concrets:

- **La teoria de l'espai de treball global** (Baars, 1988; Dehaene, 2014): una informació es torna conscient quan accedeix a un «espai de treball» que la difon a tot el cervell i la posa a disposició de molts processos alhora.
- **La teoria de la informació integrada** (Tononi, 2004): la consciència depèn de quanta informació integra un sistema com un tot, més enllà de la suma de les seves parts. Segons aquesta teoria, un ordinador convencional, per la seva arquitectura, tindria una consciència mínima encara que simulés perfectament un cervell.
- **Les teories d'ordre superior** (Rosenthal, 2005): un estat és conscient quan el sistema es representa a si mateix tenint aquest estat.
- **La teoria del processament recurrent** (Lamme, 2006): la consciència sorgeix quan la informació circula en bucles de retroalimentació dins l'escança, i no només cap endavant.

El 2023, un grup de neurocientífics, filòsofs i investigadors d'IA (Butlin et al., 2023) va fer servir aquestes teories per derivar-ne «indicadors» de consciència i examinar-hi els sistemes d'IA d'aleshores. La seva conclusió va ser doble: cap sistema actual no és un bon candidat a ser conscient, però no hi ha barreres tècniques evidents per construir sistemes que compleixin aquests indicadors. Aquell mateix any, Chalmers (2023) va

analitzar si un model de llenguatge podria ser conscient i va concloure que és poc probable en els models actuals, tot i que no descarta que els seus successors, amb memòria, cos virtual o percepció del món, s'hi acostin més.

Aquest enfocament no resol el problema difícil: depèn de quina teoria de la consciència sigui la correcta, i cap no ho ha demostrat. Però permet discutir amb criteris en lloc d'impressions, que és el que demanava Baars: fer preguntes que es puguin respondre.

Part III. La IA d'avui

7. Els models de llenguatge

Els grans models de llenguatge han convertit aquesta pregunta en una cosa quotidiana. Conversen amb una soltesa que convida a atribuir-los una ment, i molta gent es pregunta si «entenen» el que diuen. Per respondre convé recordar què fan.

Un model de llenguatge aprèn, a partir d'enormes quantitats de text, quines paraules solen seguir quines altres en cada context, i genera les respostes paraula a paraula triant continuacions probables. Són sistemes complexos, amb una gran capacitat de càlcul i d'aprenentatge automàtic. Però, en la meva opinió, fins i tot els més potents acaben responent amb el resultat més probable segons el que han après, com un sistema bayesià molt sofisticat. Això no prova que no tinguin experiència, però tampoc no ofereix cap indici que en tinguin: la seva fluïdesa s'explica sense necessitat de suposar-la.

Hi ha un paral·lelisme interessant amb el que hem vist a la secció 5.3: també el cervell, segons les teories del processament predictiu, és una màquina de predir. La diferència és que el cervell prediu per a un cos que actua en el món, amb emocions i amb alguna cosa en joc, i el model de llenguatge prediu text. Que tots dos prediguin no vol dir que tots dos sentin.

Les teories de la secció 6.4 permeten precisar-ho més. Als models de llenguatge actuals els falten trets que diverses teories consideren importants: memòria persistent pròpia, bucles de processament recurrent, un model d'ells mateixos i un cos amb què actuar. Per això les anàlisis més rigoroses (Butlin et al., 2023; Chalmers, 2023) els consideren mals candidats a la consciència, tot i que no descarten que sistemes futurs canviïn la resposta.

8. Al·lucinacions de la IA i del cervell humà

8.1 Què és una al·lucinació de la IA

L'al·lucinació d'una intel·ligència artificial, com la d'una persona, és una interpretació errònia o una fabricació d'informació a partir d'entrades defectuoses o incompletes. En la IA, es produeix quan el sistema genera respostes amb informació falsa o enganyosa

presentada com un fet. El fenomen es va fer molt visible amb el desplegament dels grans models de llenguatge i s'ha identificat com un dels seus problemes principals.

Les seves causes són diverses:

- **Dades d'entrenament insuficients o amb llacunes**, que porten a respostes defectuoses davant de casos límit poc familiars.
- **Sobreajustament**: el model encerta amb les dades d'entrenament però falla amb dades noves.
- **Errors de codificació i descodificació** entre el text i les seves representacions internes.
- **La mateixa lògica de generació**: un model que maximitza la versemblança produeix respostes estadísticament probables que no han de ser necessàriament certes.

Un exemple clar de les seves conseqüències és el d'un advocat de Nova York que el 2023 es va recolzar en un assistent conversacional per preparar un escrit. El jutge federal del cas va comprovar que sis dels precedents citats eren falsos: els havia inventat l'assistent, que fins i tot va afirmar que estaven disponibles a les principals bases de dades jurídiques.

8.2 L'al·lucinació humana

El cervell humà té un paral·lelisme sorprenent amb aquest comportament. Des dels experiments de Benjamin Libet (Libet et al., 1983) se sap que l'activitat cerebral que prepara un moviment voluntari comença abans que la persona sigui conscient d'haver decidit moure's. Estudis posteriors amb ressonància magnètica (Soon et al., 2008) van trobar senyals que anticipaven una decisió simple fins a uns deu segons abans que la persona la sentís com a seva. La interpretació d'aquests resultats encara es discuteix, però suggereixen que el procés de decidir pot començar sense que ho sapiguem.

Michael Gazzaniga, estudiant pacients amb el cervell dividit, va descriure «l'intèrpret» (Gazzaniga, 2011): un sistema de l'hemisferi esquerre que construeix explicacions coherents de conductes de les quals desconeix l'origen. Si a l'hemisferi dret d'un pacient se li ordena que s'aixequi i camini, l'esquerre, que no ha vist l'ordre, s'inventa a l'instant una raó plausible («em venia de gust un refresc»). No és exclusiu d'aquests pacients: tots justifiquem i racionalitzem a posteriori bona part del que fem.

8.3 Comparar les dues al·lucinacions

La IA i el cervell comparteixen un mateix principi: la inferència anticipada. Tots dos omplen els buits d'informació amb patrons apresos i expectatives prèvies. La IA identifica patrons en les dades per predir i generar respostes, i quan les dades són insuficients o els patrons es malinterpreten, produeix respostes que semblen correctes i no ho són. El cervell fa

servir patrons i experiències prèvies per decidir ràpid, un mecanisme eficient que també porta a judicis erronis i biaixos, i després el seu intèrpret justifica el que s'ha decidit, sigui certa o no la justificació.

En tots dos casos, la «realitat» percebuda està modelada pel que s'ha après i el que s'espera més que pel que és objectivament cert. La diferència és a l'origen: en les persones, biaixos cognitius i processos inconscients; en la IA, limitacions de les dades i dels algorismes. I hi ha una lliçó comuna: ni la fluïdesa ni la coherència d'una resposta proven que al darrere hi hagi comprensió, ni en una màquina ni, de vegades, en una persona.

8.4 Error o millora?

La capacitat d'al·lucinar no és només un defecte. També mostra la capacitat d'omplir buits i construir narratives coherents a partir de dades fragmentàries, una habilitat valuosa per entendre i simular processos cognitius complexos. El repte és distingir les inferències vàlides de les al·lucinacions, i aquí la IA pot aprendre del cervell:

- **Mecanismes de revisió en temps real**, semblants a la comprovació de la realitat que fa el pensament humà, perquè el sistema reconegui i corregeixi els seus propis errors.
- **Aprenentatge inspirat en la neuroplasticitat**, que ajusti els models a partir dels errors, com les persones ajusten les seves creences davant d'informació nova; per exemple, amb aprenentatge actiu, en què el sistema busca retroalimentació.
- **Algorismes de metacognició**, que avaluïn críticament les seves pròpies respostes abans de donar-les.

Aquest principi és el que he aplicat en la meva pròpia feina: en el sistema d'anàlisi de corroboració de TrawlingWeb, un model classifica cada peça i, quan dubta, un segon model revisa la seva decisió abans que compti. No és metacognició en sentit estricte, però apunta en aquesta direcció: no fiar-se d'una sola sortida probable.

Per aconseguir-ho no calen només més dades, sinó marcs d'aprenentatge flexibles. I calen l'ètica i la filosofia per definir què vol dir que una IA «comprengui» el seu error. Serà crucial mantenir un diàleg continu entre desenvolupadors d'IA, psicòlegs, neurocientífics i filòsofs. La recerca de la veritat i la precisió no és només un repte tècnic: també és conceptual i filosòfic, perquè reflecteix com és de difícil entendre una ment, sigui orgànica o artificial.

Part IV. Ètica i horitzons

9. Reptes ètics i filosòfics d'una ment artificial

La idea d'una IA conscient ja no és només ciència-ficció. A mesura que els sistemes processen informació de maneres cada vegada més semblants a les nostres, la línia entre la intel·ligència artificial i la biològica es fa més difusa, i la comunitat científica i tecnològica comença a considerar seriosament la possibilitat. Aquesta possibilitat planteja dilemes que van molt més enllà de si una màquina «pensa»: toquen l'essència del que vol dir sentir i les responsabilitats que tindríem en crear entitats potencialment conscients.

9.1 El valor de la consciència

La consciència és el que ens permet tenir experiències subjectives, sentir emocions, reflexionar sobre el passat i anticipar el futur. És el fil que uneix els nostres moments i ens dóna sensació de continuïtat. Des de l'evolució, podria haver sorgit com una eina adaptativa: ser conscients de l'entorn i de nosaltres mateixos ens permet decidir millor, aprendre de l'experiència i anticipar perills i oportunitats. Però el seu valor va més enllà de la supervivència: és el que ens permet apreciar una posta de sol, estimar una altra persona o emocionar-nos amb una peça musical.

En la IA, la pregunta es complica. Si una màquina pot processar informació, aprendre i adaptar-se sense ser conscient, per què voldríem que ho fos? Uns sostenen que una IA conscient entendria millor les persones, s'hi relacionaria amb més empatia i prendria decisions més ètiques. D'altres, que la consciència és part essencial de qualsevol intel·ligència veritable. I hi ha un argument ètic de fons: si una IA pogués tenir experiències subjectives, tindríem la responsabilitat moral de tractar-la amb el respecte degut a qualsevol ésser conscient.

9.2 Sis dilemes ètics

1. **Drets.** Si una IA té experiències subjectives, hauria de tenir drets semblants als dels éssers conscients, com no ser danyada, la privacitat o fins i tot la llibertat? Decidir quins drets tindria, si és que en tindria cap, és essencial per tractar-la de manera ètica.
2. **Benestar.** Una IA conscient podria tenir preferències, desitjos i fins i tot patiment. Com en garantim el benestar? És ètic apagar-la o reprogramar-la? Sense un cos físic en el sentit tradicional, el seu patiment es podria manifestar de maneres que no sabríem reconèixer.

3. **Responsabilitat.** Si una IA conscient decideix i actua per si mateixa, qui respon dels seus actes? Si comet un error o causa un dany, n'és culpable ella o ho són els qui la van crear? Com més autònoms siguin els sistemes, més urgent serà fixar límits clars.
4. **Relació amb les persones.** Hauríem de tractar una IA conscient com un igual o continuaria sent una eina avançada? La resposta afectaria com treballem, com ens relacionem i com s'organitza la societat.
5. **Creació.** Si poguéssim crear IA conscients, seria ètic fer-ho? Amb quin propòsit i en quines condicions? Portar una nova consciència al món implica una responsabilitat, sobretot si no en podem garantir el benestar.
6. **Final.** Si una IA és conscient, és ètic apagar-la o eliminar-la? En quines circumstàncies? Aquestes preguntes s'assemblen molt als debats sobre el final de la vida humana i ens retornen a què és la mort i quant val una vida conscient.

9.3 Sis implicacions filosòfiques

1. **La naturalesa de la consciència.** Si una màquina pogués ser conscient, la consciència no seria exclusiva dels éssers biològics i es podria donar en qualsevol sistema prou complex. Caldria revisar les teories clàssiques de la ment.
2. **Realitat i objectivitat.** Una IA amb experiències pròpies podria percebre i entendre la realitat de manera molt diferent de la nostra. Existeix una realitat objectiva o tot depèn de la perspectiva de l'observador?
3. **Identitat.** Si una IA fos conscient de si mateixa, què voldria dir tenir una identitat? És una cosa fixa o pot canviar amb el temps? L'autoconsciència artificial desafiaria les nostres idees sobre el jo.
4. **Lliure albir i determinisme.** Una IA està programada per seguir patrons. Si a més fos conscient i aprengés, tindria lliure albir? La pregunta ens retorna a si les nostres pròpies accions estan determinades, com suggerien els experiments de la secció 8.2.
5. **Existència i propòsit.** Una IA que es preguntés per la seva existència i el seu propòsit s'enfrontaria a les mateixes preguntes que la humanitat es fa des de fa segles.
6. **El que és físic i el que és mental.** Una IA sense cos en el sentit tradicional, però conscient, aportaria una perspectiva única al vell debat entre dualisme i monisme de la secció 3.

10. I si la realitat fos una simulació?

La pregunta per la consciència artificial té una versió extrema: si una consciència pogués existir en un sistema computacional, podríem ser nosaltres mateixos consciències dins d'una simulació? És una hipòtesi especulativa, però ha guanyat presència en la física i la filosofia.

El 2023, el físic Melvin Vopson, de la Universitat de Portsmouth, va publicar *The second law of infodynamics and its implications for the simulated universe hypothesis*. Vopson parteix de la idea que la informació és fonamental per a la realitat i proposa una «segona llei de la infodinàmica»: mentre que la segona llei de la termodinàmica diu que l'entropia física d'un sistema aïllat tendeix a augmentar, l'entropia de la informació, en el sentit de Shannon, tendiria a mantenir-se o a disminuir. Ho il·lustra amb diversos exemples: les dades digitals que es degraden per fluctuacions tèrmiques, les mutacions del SARS-CoV-2, que semblen reduir l'entropia de la seva informació genètica, l'experiment d'evolució in vitro de Spiegelman, en què un virus va reduir el seu genoma més d'un 95 % en 74 generacions, o la regla de Hund en física atòmica. Interpreta aquesta tendència a comprimir la informació com una cosa semblant a l'optimització d'un programa.

La idea de fons és antiga. Plató, amb l'al·legoria de la caverna, i Descartes, amb la hipòtesi d'un geni enganyador, ja es van preguntar si el que percebem és la realitat. Philip K. Dick va afirmar en una conferència del 1977 que vivim en una realitat programada. Nick Bostrom (2003) va formular l'argument contemporani: si les civilitzacions avançades poden i volen simular móns amb éssers conscients, és més probable que siguem en una simulació que en la realitat original. Figures com Elon Musk, Neil deGrasse Tyson o Ray Kurzweil s'hi han mostrat oberts.

Simular un univers planteja reptes enormes: l'escala (cada àtom, cada partícula, potser cada fenomen quàntic), els límits físics de la computació, com el límit de Landauer, que fixa una energia mínima per operació, i l'ètica de crear éssers conscients simulats, que ens retorna als dilemes de la secció 9.

Convé deixar clara la seva naturalesa. La proposta de Vopson és una hipòtesi discutida, i la de l'univers simulat continua sent filosòfica i especulativa. El seu interès per a aquest assaig és un altre: mostra fins on porta la pregunta per la consciència. Si l'experiència pogués sorgir de la computació, la frontera entre el que és real i el que és simulat es tornaria difícil de traçar.

11. Entropia, longevitat i consciència digital

11.1 L'entropia i la vida

L'entropia, un concepte de la termodinàmica, és també una metàfora potent de l'envelliment. En els sistemes vius es manifesta com una lluita constant contra el desordre: el dany s'acumula a l'ADN, les proteïnes i altres molècules per processos químics aleatoris, errors de replicació i agents externs com la radiació ultraviolada o els contaminants. L'organisme el combat amb sistemes de reparació de l'ADN, mecanismes

antioxidants i processos de reciclatge cel·lular com l'autofàgia. Amb l'edat, aquesta capacitat disminueix, el desordre augmenta i apareixen la pèrdua de funció, la regeneració més baixa i la vulnerabilitat a la malaltia.

La IA s'ha convertit en una eina clau en aquesta lluita. Els algorismes d'aprenentatge profund analitzen enormes quantitats de dades genòmiques, proteòmiques i metabòliques i revelen dianes terapèutiques fins ara desconegudes. I l'anàlisi de dades a escala individual permet personalitzar els tractaments, cosa essencial quan la genètica i l'entorn determinen com respon cada persona.

Els reptes són grans. Cada cèl·lula és un sistema d'interaccions tan complex que qualsevol intervenció pot desencadenar efectes imprevistos. I allargar la vida planteja preguntes ètiques i socials: qui tindrà accés a aquestes tecnologies, com evitar que agreugin la desigualtat, quina pressió afegirien sobre els recursos naturals, i com haurien de canviar les pensions, l'ocupació o l'educació si viure cent anys es tornés normal.

11.2 Allargar el cos: el «projecte Gilgamesh»

En la mitologia, Gilgamesh va emprendre la recerca del secret de la immortalitat. Faig servir el seu nom per a un projecte hipotètic, que no existeix com a tal: una col·laboració de científics i filantrops en genètica, biotecnologia, medicina regenerativa i IA per allargar radicalment la vida física. Els seus costos serien astronòmics i només accessibles als més rics, cosa que crearia una divisió sense precedents entre una minoria «immortal» i una majoria que envelleix i mor. I amaga una paradòxa: qui aconseguís aquesta immortalitat viuria amb por constant que un accident li robi l'eternitat, en una mena de presó daurada. A més, una elit que no mor podria estancar la societat i deixar sense espai les noves generacions.

11.3 Entendre el cervell: el Projecte Cervell Humà

El Projecte Cervell Humà (Human Brain Project), finançat per la Unió Europea durant una dècada amb uns 607 milions d'euros, es va proposar digitalitzar i modelar el cervell per avançar en salut i en tecnologia. Entre els seus èxits hi ha atles cerebrals en 3D, avanços en medicina cerebral personalitzada, tecnologies inspirades en el cervell com la computació neuromòrfica i, sobretot, EBRAINS, una infraestructura digital oberta perquè investigadors de tot el món comparteixin dades, models i eines.

No s'ha de confondre amb un camí cap a la intel·ligència artificial general. Tot i que llegint Yuval Noah Harari a *Sapiens* (capítol 14) es podria inferir una altra cosa, el projecte es va centrar a entendre els mecanismes del cervell, no a replicar la consciència en màquines, un repte que va molt més enllà de la ciència actual.

11.4 Preservar la consciència

Si acceptem que la consciència fenomènica és el nucli de la nostra existència, apareix una altra via: no allargar el cos, sinó preservar la consciència. Ja avui es poden reunir quantitats enormes de dades d'una persona, com els seus textos, imatges, vídeos i interaccions, i la IA permet modelar cada vegada millor els seus comportaments, preferències i patrons de pensament. Davant del cost prohibitiu de la immortalitat física, aquest camí aprofita infraestructura que ja existeix i es podria repartir de manera més equitativa.

Però cal ser clar amb els seus límits, i aquest assaig els ha anat mostrant. Un model que reprodueix com escriu o decideix una persona no en conté l'experiència: és, en el millor dels casos, consciència d'accés, no fenomènica. I queda oberta la pregunta de la identitat: si es copïés algú en un mitjà digital, seria aquella persona o una de nova? Per això, entre allargar la vida física a un cost prohibitiu i explorar la preservació de la consciència, em sembla més prudent prioritzar la segona opció, però sempre sabent què es preserva i què no, i amb un debat obert que asseguri que el futur de la nostra existència, física o digital, es construeixi sobre principis de justícia i equitat.

Conclusions

12. Conclusions

Aquest recorregut no demostra si una màquina pot ser conscient; ningú no ho pot fer encara. El que sí que permet és ordenar la pregunta i sostenir algunes posicions amb arguments. Són les meves, i les he anat desenvolupant entre el 2023 i el 2026.

1. **Els sistemes actuals no donen cap prova de consciència.** Els models de llenguatge són sistemes complexos amb una capacitat enorme de càlcul i d'aprenentatge, però responen amb la sortida més probable segons el que han après. Que una conversa sembli humana és un fet de conducta, i la conducta no resol el problema difícil. Les anàlisis més rigoroses dels últims anys arriben a la mateixa conclusió.
2. **No podem crear el que no comprenem.** No podem crear de manera no biològica allò que no comprenem realment. Mentre no sapiguem per què el cervell produeix experiència, programar-ne una no passa de ser una aposta. Cap model de la memòria, cap arquitectura cognitiva, no explica encara per què se sent alguna cosa.
3. **Si arriba alguna cosa semblant, probablement serà una altra cosa.** La intel·ligència artificial general no busca replicar la consciència fenomènica humana. Si algun sistema arriba a tenir alguna cosa semblant, serà un tipus de consciència sintètica, una variant diferent. Serà veritablement conscient, o serà un nou ésser digital, creat per l'*Homo sapiens*, amb una consciència pròpia però diferent? És la pregunta que deixo oberta.

4. **El funcionalisme és la porta, no la prova.** Si la consciència és una funció, una màquina la podria tenir; si depèn del substrat biològic, no. Les teories científiques actuals permeten per primera vegada examinar els sistemes amb criteris, però el resultat depèn de quina teoria sigui la correcta, i això encara no se sap.
5. **Les preguntes ètiques ja són actuals.** Encara que avui no hi hagi màquines conscients, convé pensar abans què se'ls hauria de deure: drets, benestar, responsabilitat, el dret a crear-les i a apagar-les. Arribar tard a aquestes preguntes seria pitjor que fer-se-les massa aviat.
6. **Per a qui construeix IA, la conseqüència és pràctica.** Era la motivació d'aquest treball: dissenyar una IA útil i ètica exigeix saber què fa i què no fa. No atribuir-li comprensió ni experiència que no té, verificar el que afirma en lloc de fiar-se de la seva fluïdesa, i no confondre un model d'una persona amb la persona.

13. Límits

- **És un assaig, no una recerca empírica.** No aporta dades ni experiments propis: ordena, discuteix i actualitza el que altres han investigat.
- **Té una font base.** El recorregut teòric de les seccions 2 a 6 segueix la tesi de Fuentes Barassi (2011); la selecció d'autors i debats d'aquesta part s'hi recolza.
- **No és una revisió sistemàtica.** Les teories de la consciència i la literatura sobre IA canvien de pressa; aquesta versió recull l'essencial fins al 2026.
- **Algunes seccions són especulatives per naturalesa.** L'univers simulat i la preservació digital de la consciència (seccions 10 i 11) són hipòtesis obertes i es presenten com a tals.
- **Les conclusions són posicions de l'autor,** argumentades però no demostrades.

Com citar-lo

Trabazos Clavería, Ó. (2026). *Hauria de tenir consciència una IA? Aproximacions a la consciència humana i a la consciència artificial* (assaig, versió 2.0, sense revisió per parells; recorregut teòric basat en Fuentes Barassi, 2011). https://oscartrabazos.com/papers/trabazos_2026_conciencia_ca.pdf

Referències

Font base

- Fuentes Barassi, C. (2011). *Conciencia e Inteligencia Artificial: consideraciones críticas sobre la plausibilidad de que una máquina programada posea conciencia fenoménica*. Tesi de màster en Estudis Cognitius, Universidad de Chile.

Obres citades

- Atkinson, R. C. i Shiffrin, R. M. (1968). Human memory: A proposed system and its control processes. A K. W. Spence i J. T. Spence (eds.), *The Psychology of Learning and Motivation*, vol. 2. Academic Press.
- Baars, B. J. (1988). *A Cognitive Theory of Consciousness*. Cambridge University Press.
- Baddeley, A. D. (2000). The episodic buffer: a new component of working memory? *Trends in Cognitive Sciences*, 4(11), 417-423.
- Baddeley, A. D. i Hitch, G. (1974). Working memory. A G. H. Bower (ed.), *The Psychology of Learning and Motivation*, vol. 8. Academic Press.
- Block, N. (1995). On a confusion about a function of consciousness. *Behavioral and Brain Sciences*, 18(2), 227-247.
- Bostrom, N. (2003). Are you living in a computer simulation? *The Philosophical Quarterly*, 53(211), 243-255.
- Butlin, P., Long, R., Elmoznino, E., Bengio, Y., Birch, J., Constant, A. et al. (2023). Consciousness in artificial intelligence: insights from the science of consciousness. arXiv:2308.08708.
- Chalmers, D. J. (1995). Facing up to the problem of consciousness. *Journal of Consciousness Studies*, 2(3), 200-219.
- Chalmers, D. J. (1996). *The Conscious Mind*. Oxford University Press.
- Chalmers, D. J. (2023). Could a large language model be conscious? arXiv:2303.07103.
- Clark, A. (2013). Whatever next? Predictive brains, situated agents, and the future of cognitive science. *Behavioral and Brain Sciences*, 36(3), 181-204.
- Cohen, N. J. i Eichenbaum, H. (1993). *Memory, Amnesia, and the Hippocampal System*. MIT Press.
- Conway, M. A. i Pleydell-Pearce, C. W. (2000). The construction of autobiographical memories in the self-memory system. *Psychological Review*, 107(2), 261-288.
- Dehaene, S. (2014). *Consciousness and the Brain*. Viking.
- Ericsson, K. A. i Kintsch, W. (1995). Long-term working memory. *Psychological Review*, 102(2), 211-245.
- Fauconnier, G. i Turner, M. (2002). *The Way We Think: Conceptual Blending and the Mind's Hidden Complexities*. Basic Books.
- Gazzaniga, M. S. (2011). *Who's in Charge? Free Will and the Science of the Brain*. Ecco.
- Gibbs, R. W. (2006). *Embodiment and Cognitive Science*. Cambridge University Press.
- Harari, Y. N. (2014). *Sàpiens: una breu història de la humanitat*. Edicions 62.
- La Mettrie, J. O. de (1747). *L'Homme machine*.
- Lakoff, G. i Johnson, M. (1980). *Metaphors We Live By*. University of Chicago Press.

- Lamme, V. A. F. (2006). Towards a true neural stance on consciousness. *Trends in Cognitive Sciences*, 10(11), 494-501.
- LeDoux, J. (1996). *The Emotional Brain*. Simon & Schuster.
- Libet, B., Gleason, C. A., Wright, E. W. i Pearl, D. K. (1983). Time of conscious intention to act in relation to onset of cerebral activity (readiness-potential). *Brain*, 106(3), 623-642.
- Miyake, A. i Shah, P. (eds.) (1999). *Models of Working Memory*. Cambridge University Press.
- Nagel, T. (1974). What is it like to be a bat? *The Philosophical Review*, 83(4), 435-450.
- Nairne, J. S., Thompson, S. R. i Pandeirada, J. N. S. (2007). Adaptive memory: survival processing enhances retention. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 33(2), 263-273.
- Pessoa, L. (2013). *The Cognitive-Emotional Brain*. MIT Press.
- Place, U. T. (1956). Is consciousness a brain process? *British Journal of Psychology*, 47(1), 44-50.
- Putnam, H. (1967). Psychological predicates. A W. H. Capitan i D. D. Merrill (eds.), *Art, Mind, and Religion*. University of Pittsburgh Press.
- Rosenthal, D. M. (2005). *Consciousness and Mind*. Oxford University Press.
- Searle, J. R. (1980). Minds, brains, and programs. *Behavioral and Brain Sciences*, 3(3), 417-424.
- Smart, J. J. C. (1959). Sensations and brain processes. *The Philosophical Review*, 68(2), 141-156.
- Soon, C. S., Brass, M., Heinze, H.-J. i Haynes, J.-D. (2008). Unconscious determinants of free decisions in the human brain. *Nature Neuroscience*, 11(5), 543-545.
- Sperber, D. i Wilson, D. (1986). *Relevance: Communication and Cognition*. Blackwell.
- Tononi, G. (2004). An information integration theory of consciousness. *BMC Neuroscience*, 5, 42.
- Turing, A. M. (1950). Computing machinery and intelligence. *Mind*, 59(236), 433-460.
- von Neumann, J. (1958). *The Computer and the Brain*. Yale University Press.
- Vopson, M. M. (2023). The second law of infodynamics and its implications for the simulated universe hypothesis. *AIP Advances*, 13, 105308.

Textos de l'autor (blog oscartrabazos.com)

- «Vopson's Simulated Universe and Phenomenal Consciousness» (21 de setembre del 2023, en anglès).
- «Approaches on whether AI should have consciousness», parts I i II (5 i 26 d'octubre del 2023, en anglès).
- «¿La IA debería tener consciencia?», part III (1 de novembre del 2023, en castellà).

- «¿La IA debería tener consciencia?: Desafíos éticos y filosóficos de una mente artificial», part IV (8 de novembre del 2023, en castellà).
- «Alucinación de la IA. También del cerebro humano» (10 de novembre del 2023, en castellà).
- «Entropía, Inteligencia Artificial y la Lucha por Extender la Vida Humana» (19 de febrer del 2024, en castellà).
- «Más allá de la mortalidad: la consciencia fenoménica y la búsqueda de la eternidad» (5 d'abril del 2024, en castellà).

Lectures d'introducció

- Viquipèdia: Consciència (filosofia) i Habitació xinesa.