Intel·ligència artificial, creences, consciència i solidaritat humana

Sara Lumbreras Sancho
Universitat Pontifícia Comillas – Instituto de Investigación Tecnológica.

Introducció: la importància renovada de l'estudi de les creences

Les creences són subjacents al comportament moral en diversos sentits. Primerament, són les creences les que defineixen els valors que ens guien i la seva importància relativa, així com les normes a què voluntàriament ens adherim. Tot i la seva importància, encara ens falta molt per saber, dels processos associats a les creences.

Saber què són (què creiem i de quina manera), com es formen i canvien, és crucial per entendre des de l'adopció de tecnologies (com la intel·ligència artificial) fins a les conversions religioses, passant per qüestions com les actituds polítiques o les pràctiques de salut. A més, els processos de creença també tenen, com veurem, importants implicacions pràctiques en el desenvolupament i l'aplicació de sistemes

intel·ligents. Reconèixer les diferències entre els processos humans de creure i els que podem simular a la màquina ens ajuden a dissenyar interaccions no només més eficients sinó més ètiques. Revisarem la Llei d'Intel·ligència Artificial europea en aquest context per il·luminar algunes de les implicacions més importants.

Creences i consciència

Endinsar-se en la naturalesa de les creences i el seu tractament computacional no només enriqueix la nostra comprensió del funcionament intern del cervell humà, sinó que obre camins cap a la creació d'intel·ligència artificial més avançada i, potser, més «humana».

Endinsar-se en la naturalesa de les creences i el seu tractament computacional obre camins per la creació d'intel·ligència artificial més «humana».

Les ciències cognitives han intentat oferir models de com funciona la ment humana. Una línia predominant d'aquests models és l'enfocament representacional, on les idees i l'accés mental a la realitat s'entenen com a representacions de la realitat (operacions matemàtiques en sentit ampli). Tot i això, els models computacionals no aconsegueixen descriure completament la complexitat dels fenòmens mentals, per exemple, la riquesa de l'experiència subjectiva.

Això es reflecteix de manera molt interessant en algunes teories recents sobre la consciència, que destaquen els aspectes de la ment que van més enllà de la mera computació. Un exemple és el dels robots amb sensors i actuadors que han elaborat models de si mateixos a partir d'informació externa, descrivint-se com a sistemes encarnats i integrats.[1] No obstant això, no és clar que la intel·ligència artificial encarnada proporcioni un model convincent per representar la ment i la cognició humanes.[2] La forma més sofisticada

[1] M. HOFFMANN – R. PFEIFER, «Robots as powerful allies for the study of embodied cognition from the bottom up», dins A. NEWEN - L. DE BRUIN - S. GALLAGHER (ed.), *The Oxford Handbook of 4e Cognition*, Oxford, Oxford University Press, 2018, p. 841-862.

[2] R. MANZOTTI, «Embodied AI beyond embodied cognition and enactivism», *Philosophies* 4(3) (2019), p. 39.

d'aquestes visions és, probablement, la teoria 4E d'Oxford (ment encarnada, integrada, activa i estesa): la ment encarnada subratlla com el nostre cos influeix en el pensament; la integrada, com l'entorn afecta la nostra cognició; l'activa veu la cognició com a resultat de la interacció amb l'entorn; i l'estesa considera que objectes i tecnologies externes són part dels nostres processos cognitius, i desafia la idea que la ment opera només internament en el cervell, perquè totes les eines que utilitzem també en formen part.[3]

Els desenvolupaments recents en la robòtica i la teoria de la informació han enriquit aquestes perspectives, augmentant així la pluralitat d'interpretacions sobre la consciència. Un exemple d'això són les teories que intenten definir la consciència com una mesura d'integració d'informació.[4] Aquestes teories han rebut crítiques intenses, perquè impliquen el panpsiquisme: qualsevol processament d'informació tindria, en principi, un cert grau de consciència, segons aquestes teories.

El procés de creure posa a prova els models cognitius més sofisticats. Creure és una experiència humana comuna, des de creences simples com «està plovent ara mateix» fins a les més abstractes i compromeses, com «crec que la meva vida té sentit». Tradicionalment, el procés de creure ha ocupat un lloc secundari en comparació amb actituds cognitives «superiors», a causa de la probabilitat de veritat de les creences, que no necessàriament són veritables sinó que varien en graus de força, certesa i confiança.[5]

El fet que la veritat de les creences s'expressi en termes de probabilitat és crucial per comprendre com sorgeixen, s'estabilitzen i fins i tot desapareixen les creences. Des dels seus inicis, la intel·ligència artificial ha ofert models que presenten estimacions probabilístiques per al valor de veritat de les

[3] A. NEWEN – L. DE BRUIN – S. Gallagher (ed.), *The Oxford Handbook of 4e Cognition*, Oxford, Oxford University Press, 2018.

[4] F. TONONI *et alii*, «Integrated Information Theory: From Consciousness to Its Physical Substrate», *Nature Reviews Neuroscience* 17 (2016), p. 450-461.

[5] M. AYERS – M. R. ANTOGNAZZA, «Knowledge and belief from Plato to Locke», dins M. AYERS (ed.), *Knowing and Seeing: Groundwork for a New Empiricism*, Oxford, Oxford University Press, 2019. L. ERIKSSON – A. HÁJEK, «What Are Degrees of Belief?», *Studia Logica* 86(2) (2007), p. 183-213. F. HUBER, «Belief and Degrees of Belief», dins F. HUBER - C. SMITHPETRI (ed.), *Degrees of Belief*, Berlin/Heidelberg, Springer, 2009, p. 133.

creences. Els motors d'inferència (sistemes que apliquen lògica per deduir conclusions a partir de premisses donades) i les xarxes bayesianes (models estadístics que representen variables i les seves dependències condicionals a través de grafs dirigits) són eines fonamentals per a la intel·ligència artificial per modelar i raonar sota incertesa.[6] Hi ha extensions dels algorismes d'intel·ligència artificial fonamentals que tenen en compte la incertesa.

A més, alguns desenvolupaments relativament recents de la intel·ligència artificial, com l'aprenentatge per reforç, han augmentat les expectatives que podria ajudar-nos a entendre millor el funcionament de les nostres ments i, per tant, a omplir alguns buits en els models computacionals actuals per als processos cognitius, incloent-hi la formació i la constitució de creences. Sistemes d'intel·ligència artificial basats en el reconeixement de patrons poden dur a terme tasques que s'acosten a alguns aspectes del procés de creure, com ara creure alguna cosa a conseqüència d'un patró recurrent d'esdeveniments o creure com a resultat de l'aprenentatge a partir de dades noves. Això podria suggerir que els sistemes d'intel·ligència artificial funcionarien de manera similar a alguns processos mentals humans, a l'hora de discernir sobre la formació, el desenvolupament, la confirmació o la negació de creences.

Així, hi ha una relació bidireccional entre l'estudi de les creences i la intel·ligència artificial: d'una banda, l'estudi de les creences ens pot inspirar a crear sistemes més humans i, de l'altra, la intel·ligència artificial pot proporcionar models més sofisticats per comprendre els processos relacionats amb les creences.

La creença com a procés cognitiu

Un procés cognitiu és qualsevol funció mental que participa en l'acte d'adquirir coneixement i comprensió a través del pensament, l'experiència i els sentits. Aquests processos inclouen una àmplia gamma d'operacions mentals que ens permeten interpretar, processar i emmagatzemar informació

[6] F. V. JENSEN, *An Introduction to Bayesian Networks*, London, UCL Press, 1996.

La intel·ligència artificial pot proporcionar models més sofisticats per comprendre els processos relacionats amb les creences.

essencial per a la presa de decisions, la solució de problemes, l'aprenentatge i la comprensió del món. Exemples de processos cognitius són la percepció, que ens permet interpretar i donar sentit als estímuls sensorials; la memòria, l'emmagatzemament i la recuperació d'informació; l'aprenentatge, l'adquisició de coneixements o habilitats nous; el raonament, la capacitat de resoldre problemes i prendre decisions; i el llenguatge, la nostra eina per comunicar pensaments i idees.

El procés de creure també és un procés cognitiu, i és crític perquè les creences configuren fonamentalment el comportament individual en establir valors i permetre prediccions d'esdeveniments futurs.[7] Tot i la seva importància, l'exploració científica de les creences ha estat històricament escassa, en part pel fet que han estat sobretot associades a l'espiritualitat,[8] si bé les creences poden tenir, com dèiem més amunt, objectes molt diferents. Només recentment s'han començat a estudiar de manera científica els processos de creure, amb un interès creixent de diverses disciplines científiques, incloent-hi la biologia evolutiva, la neurociència cognitiva, la psicologia i la psiquiatria.[9] Aquest ressorgiment està sostingut per l'evidència de la neuroimatge, que revela els correlats neuronals subjacents de l'acte de creure.[10]

Hi ha una influència doble de les creences a les nostres accions i viceversa.[11] Sacks i Hirsch van postular que les persones tendeixen a acceptar una cosa com a realitat fins que es demostra el contrari i que la formació de creences es pot entendre com el resultat del processament d'informació perceptiva i afectiva.[12] Sostenint aquesta noció, s'ha mostrat

[7] H. S. ANGEL – R. J. SEITZ, «Process of believing as fundamental brain function: The concept of Credition», *SFU Research Bulletin* 1(1) (2016), p. 1-20.

[8] H. F. ANGEL *et alii* (ed.), *Processes of believing: The acquisition, maintenance, and change in creditions*, Basel, Springer International Publishing AG, 2017.

[9] H. F. ANGEL *et alii* (ed.), *Processes of believing...*, *op. cit*.

[10] I. CRISTOFORI – J. GRAFMAN, «Neural underpinnings of the human belief system», dins H. F. Angel *et alii* (ed.), *Processes of believing...*, *op. cit*.

[11] R. F. PALOUTZIAN – K. Mukai, «Believing, remembering, and imaging: The roots and fruits of meanings made and remade», dins H. F. Angel *et alii* (ed.), *Processes of believing...*, *op. cit*, p. 39-49.

[12] O. SACKS – J. HIRSCH, «A neurology of belief», *Annals of Neurology* 63(2)

que la integració de la cognició i l'emoció té lloc a l'escorça prefrontal lateral.[13]

En aquest article destaquem el model *Credition* de les creences, en què es diferencien quatre paràmetres característics per a cada creença: proposició, certesa, emoció i poder.[14] La proposició representa el contingut de la declaració (per exemple, «L'entrevista de feina ha anat bé»); la certesa reflecteix la inclinació de la persona a creure en la proposició, que també podem interpretar com la probabilitat de veritat de la creença; l'emoció reflecteix la valència afectiva de la proposició per a una persona (per exemple, es pot associar amb esperança, por, ràbia, etc.); i el poder reflecteix el grau de rellevància de la proposició, que també s'associa amb la seva intensitat afectiva. Aquest model assumeix, a més, que els processos de creure estan influenciats tant per l'individu en si mateix com per factors socioculturals i circumstàncies externes.[15] Aquest model es combina amb altres per comprendre la formació i el canvi de les creences. L'exemple més complet seria la seva associació amb la teoria de sistemes complexos,[16] que planteja com les creences estan relacionades les unes amb les altres i aquestes relacions en determinen l'evolució dinàmica.

Com dèiem, les creences guien el nostre comportament en dos sentits diferents: d'una banda, expressen els valors que es fan servir per definir els objectius de les nostres accions; d'una altra, també expressen la informació de què disposa un individu, en un moment determinat, per valorar les conseqüències de les seves accions, que són necessàriament incertes i que porten un valor emocional associat. Cal entendre la relació entre totes dues com a dinàmica: hi ha un aprenentatge associat a cada acció, que refina les creences

(2008), p. 129-130.

13 J. R. GRAY – T. S. BRAVER – M. E. RAICHLE, «Integration of emotion and cognition in the lateral prefrontal cortex», *Proceedings of the National Academy of Sciences* 99(6) (2002), p. 4.115-4.120.

14 H. S. ANGEL – R. J. SEITZ, «Process of believing as fundamental brain function: The concept of Credition», *art. cit.*

15 R. J. SEITZ – R. F. PALOUTZIAN – H. F. ANGEL, «From Believing to Belief: A General Theoretical Model», *Journal of Cognitive Neuroscience* 30(9) (2018), p. 1.254-1.264.

16 A. VESTRUCCI – S. LUMBRERAS – L. OVIEDO, «Can AI Help Us to Understand Belief?: Sources, Advances, Limits, and Future Directions»: <https://ja.cat/235ZV>.

que defineixen la nostra visió del món i que marquen quina serà l'acció que es prendrà després.

Aproximacions ètiques a la intel·ligència artificial

En els darrers anys hem estat testimonis de la creació de guies ètiques per a la intel·ligència artificial, primer basades en principis i després materialitzant-se en legislació. La proposta més ambiciosa en aquest sentit és sens dubte la Llei d'Intel·ligència Artificial de la Unió Europea. Ha suposat una proposta legislativa integral dissenyada per regular el desenvolupament, la implementació i l'ús de la intel·ligència artificial dins dels seus Estats membres (s'ha d'exigir també a qualsevol empresa amb què s'estableixin relacions comercials). Defineix un marc normatiu que garanteixi la seguretat i els drets fonamentals dels ciutadans davant dels riscos que pot suposar la intel·ligència artificial, alhora que promou la innovació i l'adopció d'aquesta tecnologia en diversos sectors.

La seva idea principal és classificar els sistemes d'intel·ligència artificial segons el nivell de risc i adaptar-hi les seves exigències. Hi ha una classificació de risc inacceptable, que prohibeix directament la realització d'algunes aplicacions. En són exemples els sistemes de rànquing social, en què els ciutadans reben puntuacions basades en el seu comportament, estabilitat financera o les opinions d'altres persones (com a curiositat, aquest tipus de sistemes ja s'han implementat a la Xina). El risc va baixant fins al nivell mínim (per exemple la intel·ligència artificial integrada a la càmera d'un telèfon mòbil que la fa servir per enfocar), que no estarien subjectes a regulació.

Les exigències principals de la Llei d'Intel·ligència Artificial són la transparència, la supervisió i la responsabilitat: la transparència és l'exigència de claredat sobre com funcionen els sistemes intel·ligents i com prenen decisions; la supervisió implica un monitoratge constant a càrrec d'entitats designades especialment per això per tal d'assegurar el compliment de les normes; i la responsabilitat garanteix que els desenvolupadors i usuaris d'intel·ligència artificial retin comptes per l'impacte i el funcionament dels seus sistemes, assegurant

la protecció dels drets individuals i la mitigació de riscos.

És útil recordar que aquest tipus d'aproximacions (basades en principis) no són les úniques que s'han discutit per implementar una ètica de la intel·ligència artificial. En aquestes converses s'ha reflectit una gran diversitat d'acords sobre l'ètica en conjunt, que s'emmarquen en dues dicotomies principals que classifiquen els enfocaments ètics segons l'objecte: l'ètica negativa i l'ètica positiva. A més, és essencial considerar com es desenvolupen aquestes perspectives, ja sigui des d'un enfocament de dalt a baix (*top down*) o de baix a dalt (*bottom up*). La combinació d'aquestes dicotomies ofereix un marc robust per analitzar les aproximacions ètiques aplicades a la intel·ligència artificial.

La primera dicotomia és la de l'ètica negativa *vs.* ètica positiva. L'ètica negativa se centra a prevenir el mal a altres éssers. En general, els codis morals que es troben en aquesta categoria solen consistir en accions que cal evitar, com ara matar, robar o mentir.[17] Aquest enfocament subratlla la importància d'establir límits a les accions de la intel·ligència artificial. En contrast, l'ètica positiva s'enfoca a crear el màxim bé possible en comptes de simplement evitar el dany. Aquests enfocaments solen ser conseqüencialistes, que avaluen el bé associat a una decisió per determinar si és el millor curs d'acció.[18]

Els enfocaments de dalt a baix i de baix a dalt completen aquesta descripció. L'ètica de dalt a baix concep les regles morals o la definició del bé ètic com a objectiu que és acceptat per l'agent moral. L'ètica kantiana és un exemple d'aquest tipus d'enfocament.[19] En el context de la intel·ligència artificial, això es traduiria en sistemes dissenyats amb principis ètics predefinits que en guien el comportament. L'ètica

[17] R. A. HOWARD – C. D. KORVER, *Ethics for the Real World: Creating a Personal Code to Guide Decisions in Work and Life*, Cambridge, Harvard Business Press, 2008.

[18] M. M. HANDELSMAN – S. KNAPP - M. C. GOTTLIEB, «Positive ethics: Themes and variations», dins *In Oxford Handbook of Positive Psychology*, Oxford, Oxford University Press, 2009, p. 105-113. F. JACKSON, «Decisiontheoretic consequentialism and the nearest and dearest objection», *Ethics* 101(3) (1991), p. 461-482.

[19] C. ALLEN – I. SMIT – W. WALLACH, «Artificial morality: Topdown, bottomup, and hybrid approaches», *Ethics and Information Technology* 7(3) (2005), p. 149-155. I. KANT – T. K. ABBOTT, *Critique of Practical Reason*, Miami, Courier Corporation, 2004.

de baix a dalt considera que és el subjecte qui selecciona els valors que guien el seu comportament i els refina progressivament en un procés d'aprenentatge que depèn de l'experiència.[20] Aquest enfocament suggereix que les intel·ligències artificials podrien desenvolupar i adaptar els principis ètics a través de la interacció amb l'entorn i l'experiència.

En el cas de la Llei d'Intel·ligència Artificial ens trobaríem amb un enfocament de dalt a baix que en realitat és negatiu, perquè més que portar-nos a maximitzar un potencial positiu, el que fa és limitar què es pot fer (això és, en el cas d'una activitat d'alt risc es limita l'activitat en si mateixa, i en el cas d'un risc mitjà es limita fer-la de manera opaca, per exemple).

Les limitacions de la màquina

Aquestes aproximacions, però, no capten la complexitat de la moral humana. Les ètiques conseqüencialistes són molt atractives més enllà dels enfocaments materialistes. No només es poden fer servir per formalitzar una maximització de beneficis sinó, des d'un punt de vista humanista, la maximització d'un bé comú, o des d'una perspectiva religiosa, contribuir de la millor manera possible al projecte diví. No obstant això, ¿aquestes serien veritablement possibles en el context de la intel·ligència artificial o són les aproximacions negatives les úniques possibles?

Hom pot entendre les limitacions de la màquina des de la seva participació només parcial en la riquesa dels processos de creure en l'ésser humà. Un primer problema és com en resulta de complicat calcular totes les conseqüències d'una acció, però el desafiament principal és definir què és realment valuós o bo, cosa que no sempre és clara i, en qualsevol cas, s'han de poder expressar en el llenguatge de programació de la intel·ligència artificial. Per exemple, és relativament fàcil programar una intel·ligència artificial per maximitzar guanys a la borsa o millorar la satisfacció del client en un

[20] R. L. CAMPBELL – J. C. CHRISTOPHER – M. H. BICKHARD, «Self and values: An interactivist foundation for moral development», *Theory & Psychology* 12(6) (2002), p. 795-823. S. A. HARDY – G. CARLO, «Moral identity: What is it, how does it develop, and is it linked to moral action?», *Child Development Perspectives* 5(3), (2011), p. 212-218.

hotel definint clarament què es considera un resultat beneficiós (per exemple, mitjançant els resultats d'una enquesta). Tanmateix, és difícilment concebible definir de manera objectivament calculable conceptes com la felicitat o la justícia. A més d'això, el model *Creditions* il·lumina que no només l'objecte de pensament és important, sinó que també és clau l'emoció que hi va associada, i aquesta és fora de les possibilitats de les màquines.

És útil recordar aquí la distinció entre llenguatges públics, com les matemàtiques o el codi de programació, que són objectius i clars per a tothom, i els llenguatges personals, que inclouen conceptes subjectius i experiències.[21] Conceptes com les emocions o valors com ara la justícia o la bellesa pertanyen a aquest àmbit personal i el seu grau de subjectivitat varia de manera notable. El que és completament objectiu i pot ser mesurat pertany al domini públic, i els conceptes personals poden perdre matisos en ser traduïts a aquest domini. Per exemple, hom pot projectar la idea d'evitar danys en termes clars i mesurables per a una tasca específica, com evitar col·lisions en un vehicle autònom, però això no captura aquesta idea de manera íntegra.

És difícilment concebible definir de manera objectivament calculable conceptes com la felicitat o la justícia.

A més, definir un bon propòsit no és suficient. Les màquines aprenen estratègies que poden portar a resultats no desitjats, com ara manipular les dades per millorar les enquestes de satisfacció del client, mostrant la necessitat addicional d'un enfocament ètic negatiu que previngui accions inacceptables. Aquest problema ha motivat la tendència a fer intel·ligència artificial «compatible amb l'ésser humà»,[22] que emfatitza que perquè un sistema d'intel·ligència artificial sigui segur no se li pot donar una definició estàtica dels seus objectius, ja que aquests es poden pervertir. Nick Bostrom mostrava que fins i tot finalitats aparentment inofensives com ara calcular les xifres del nombre pi, aviat podria convertir-se en una amenaça *existencial* per a la humanitat si la màquina

[21] J. LEACH, «Los lenguajes de la inteligencia artificial, los lenguajes de la metafísica y los lenguajes de la fe», *Scientia et Fides* 2(1) (2014), p. 81-98.

[22] S. RUSSELL, *Human compatible: AI and the problem of control*, London, Penguin, 2019.

decidia posar tots els recursos del planeta a disposició del càlcul.[23] La proposta de Russell és deixar sempre oberta la definició de metes, perquè sigui l'ésser humà qui, en contínua interacció amb la màquina, decideixi quan cal adaptar-ne els objectius. Aquest principi apareix a la Llei d'Intel·ligència Artificial com la necessitat de monitorització contínua.

Alguns autors tecnooptimistes argumenten que, en el futur, les intel·ligències artificials no només podrien actuar d'acord amb valors humans, sinó fins i tot experimentar aquests valors, desenvolupant una forma de consciència o ètica autèntica, com planteja Kurzweil. Aquest autor també suggereix que es podria crear una ment artificial, que de fet seria un sistema de reconeixement de patrons complementada amb un mòdul de valors que dotaria la màquina d'objectius i experiència moral. No obstant això, mai no proporciona idees de com es podria fer això; fins ara, a aquests enfocaments de baix a dalt els manca suport teòric sobre com sorgiria realment aquesta consciència o subjectivitat en les intel·ligències artificials. Sense aquesta experiència subjectiva, però, l'ètica de baix a dalt no és possible per a les màquines. Hem de centrar-nos en aproximacions de dalt a baix en què es plantegin externament les regles (ètica negativa) i els valors que el sistema ha d'optimitzar (ètica positiva).

Ètica i subjectivitat: empatia, solidaritat i altruisme

La relació entre l'ètica i el comportament prosocial és clau, ja que l'ètica proporciona el marc moral i els principis que sovint motiven o justifiquen la realització d'accions prosocials. En aquest sentit, el comportament prosocial es veu com la manifestació pràctica dels principis ètics; és a dir, l'aplicació de conceptes com l'altruisme, la responsabilitat pels altres i el desig de contribuir positivament a la societat. Les accions prosocials són, per tant, expressions concretes de valors ètics que promouen el benestar dels altres.

L'altruisme es defineix com la motivació per millorar el benestar d'una altra persona, en contrast amb l'egoisme, que

[23] N. BOSTROM, *Superintelligence: Paths, dangers, strategies*, Oxford, Oxford University Press, 2014.

cerca el benefici propi. La distinció entre comportament prosocial i altruisme se centra en la motivació que hi ha rere les accions beneficioses cap als altres: mentre que l'altruisme es defineix per una motivació genuïna d'ajudar els altres, el comportament prosocial abasta una gamma més àmplia d'accions beneficioses, independentment de la motivació subjacent. Per exemple, una persona pot donar a una organització benèfica pel desig genuí d'ajudar les persones necessitades (altruisme), mentre que una altra persona ho pot fer per rebre una deducció fiscal o millorar la seva imatge pública (comportament prosocial no altruista). S'ha estudiat que l'altruisme es basa en l'ètica i la filosofia (o, més aviat, de les seves creences associades), i forma el nucli no només de l'educació ètica sinó de tota educació. Batson i Powell destaquen aquesta distinció per emfatitzar que la conducta humana és multifacètica i que entendre les motivacions que hi ha rere les accions és crucial per apreciar la diversitat del comportament prosocial.

A més, l'empatia té un paper crucial en l'altruisme: l'emoció empàtica evoca una motivació veritablement altruista, amb l'objectiu últim de beneficiar la persona per la qual se sent empatia. Això desafia la suposició que fan els sectors més reduccionistes de l'estudi del comportament humà que tota acció intencionada, incloent-hi les destinades a beneficiar-ne d'altres, és egoista.

La solidaritat o comportament solidari és un altre aspecte crucial del comportament prosocial, sovint utilitzat en sociologia per referir-se a accions que beneficien el grup en conjunt. Aquest terme està estretament relacionat amb el comportament cooperatiu i assenyala una inclinació cap al benestar col·lectiu per sobre de l'individual.

Tant l'empatia com la solidaritat es basen en sentiments: l'empatia, en el fet de compartir el patiment de l'altre; la solidaritat, en l'experiència d'una identitat compartida que impulsa a cooperar davant les dificultats. La màquina, mentre no es materialitzin les profecies dels tecnooptimistes, no en pot participar. Tanmateix, sí que pot exhibir comportaments que qualificaríem de prosocials. Hi ha bots de conversa «empàtics» que emulen l'estat d'ànim del seu interlocutor humà (seleccionant les paraules més adequades a la informació que reben). Fins i tot ja existeixen aplicacions d'intel·ligència artificial que proporcionen companyia artificial

i que són valorades pels seus usuaris com a més amables i empàtiques que els éssers humans reals. Per exemple, Xiao Ice és un servei amb centenars de milers d'usuaris a la Xina que proporciona companyia artificial amb la forma d'una «xicota perfecta»[24] i es defineix comercialment com un bot de conversa empàtic. Molts dels seus usuaris defensen que els resulta més estimulant que l'amistat amb un ésser humà real, ja que el bot de conversa sempre està disponible quan se'l necessita i no exigeix res a l'usuari. Amb tot, ¿què busquem quan ens relacionem amb un altre?, ¿una connexió autèntica o un mirall de nosaltres mateixos? No ens hauríem de comportar de la mateixa manera amb una persona que amb una màquina. Tot i que probablement és desitjable no perdre mai la cordialitat amb les màquines (més per nosaltres mateixos que, òbviament, per la màquina com a tal), compartir la nostra vulnerabilitat amb un bot de conversa resulta perillós, no només en termes de privadesa sinó per la confusió de relacions inautèntiques a què ens porta. La Llei d'Intel·ligència Artificial exigeix que s'informi els consumidors si interactuen amb un bot de conversa, perquè puguin ajustar el seu comportament en conseqüència. Cal que refinem la nostra manera de relacionar-nos amb les màquines per evitar confusions perilloses.

Conclusions

Les interaccions entre la intel·ligència artificial, les creences, l'ètica i la consciència humanes són complexes i s'entrellacen profundament en el desenvolupament i l'aplicació de sistemes intel·ligents. La discussió sobre la Llei d'Intel·ligència Artificial de la Unió Europea il·lumina les implicacions ètiques i pràctiques d'aquestes tecnologies, emfatitzant com una comprensió profunda de les creences i el seu procés crea polítiques més efectives i justes.

D'altra banda, el concepte de solidaritat humana i la capacitat d'empatia i d'altruisme ressalten àrees on la intel·ligència artificial, en el seu estat actual, no pot replicar genuïnament

[24] L. ZHOU *et alii*, «The design and implementation of xiaoice, an empathetic social chatbot», *Computational Linguistics* 46 (1) (2020), p. 53-93.

l'experiència humana. Tot i que la intel·ligència artificial pot exhibir comportaments que simulen l'empatia o el suport emocional, aquestes interaccions no tenen la profunditat i l'autenticitat de les relacions humanes, cosa que planteja preguntes sobre la naturalesa de les nostres interaccions amb la tecnologia i el valor que assignem a l'autenticitat i a la connexió humana.

La intel·ligència artificial pot exhibir comportaments que simulen empatia o suport emocional, però sense la profunditat i l'autenticitat de les relacions humanes.

Cal continuar explorant la relació entre la intel·ligència artificial i els processos cognitius humans. A mesura que avancem en el desenvolupament de sistemes intel·ligents, serà imperatiu mantenir un enfocament crític i reflexiu sobre com aquestes tecnologies s'integren a la societat, impacten en les nostres interaccions i reflecteixen els nostres valors. En fer-ho, aspirem no només a avançar en el camp de la intel·ligència artificial, sinó també a aprofundir en la comprensió de la nostra pròpia naturalesa i a promoure una societat més humana. Ω

Altres referències

FRISTON, K. J. – PARR, T. – DE VRIES, B., «The graphical brain: Belief propagation and active inference», *Network Neuroscience* 1(4) (2017), p. 381-414.

GOEL, V. – DOLAN, R. J., «Explaining modulation of reasoning by belief», *Cognition* 87 (2003), B11-B22.

KURZWEIL, R., *How to Create a Mind: The Secret of Human Thought Revealed*, London, Penguin, 2012.

SCHULZ, M., «Strong Knowledge, Weak Belief? Synthese» (2021): <https://ja.cat/Lf1Di>.

SEITZ, R. J. – ANGEL, H. F., «Processes of believing a review and conceptual account», *Review of Neuroscience* 23(4) (2012), p. 303-309.

SUGIURA, M. – SEITZ, R. J. – ANGEL, H. F., «Models and neural bases of the believing process», *Journal of Behavioral and Brain Sciences* 5 (2015), p. 12-23.

WILLIAMSON, T., *Knowledge and Its Limits*, Oxford, Oxford University Press, 2002.