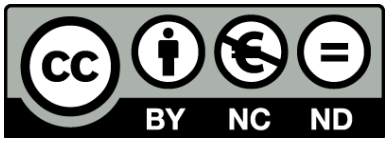
UNIVERSITAT DE BARCELONA

Reflexions ontològiques, epistemològiques i ètiques sobre la intel·ligència artificial

Pau Valls Murtra



UNIVERSITAT DE BARCELONA

Tesi doctoral

REFLEXIONS ONTOLÒGIQUES EPISTEMOLÒGIQUES I ÈTIQUES SOBRE LA INTEL·LIGÈNCIA ARTIFICIAL

Doctorand
Pau Valls Murtra

Directora i tutora
Dra. Begoña Román Maestre

Programa de Doctorat en Filosofia Contemporània i Estudis Clàssics

Facultat de Filosofia
Setembre, 2024

«…cap època ha produït mites de l'intel·lecte
de forma tan àgil com la nostra, que produeix mites
justament per l'afany d'exterminar tots els mites»

Søren Kierkegaard, *El concepte de l'angoixa*

Agraïments

Després de sis llargs anys pensant, discutint i escrivint els continguts d'aquesta tesi doctoral, em semblaria deshonest culminar l'aventura sense agrair i reconèixer les contribucions de les persones que m'han ajudat a arribar a bon port. Escric aquestes paraules després d'inacabables dies i nits intentant trobar la manera apropiada d'enllestir la present recerca —com si la recerca s'acabés algun dia—, però si aquest treball ha estat possible és gràcies a un procés continu que involucra a persones molt diverses i que, de ben segur, alguna oblidaré de mencionar.

Aquests dies no he pogut deixar de pensar en les paraules de Newton «si he vist més lluny, ha estat posant-me sobre espatlles de gegants» i l'agraïment que sento cap a els que han reflexionat abans que jo sobre els temes tan diversos que, entre molts altres, poden relacionar-se amb aquest treball: el coneixement es construeix entre tots. Tanmateix, Nietzsche deia que per més que una persona contempli les altures més sublims, només les pot reduir al seu nivell de comprensió i no entendre la profunditat del que està veient.

El meu agraïment és per totes les persones que m'han ensenyat a comprendre el que veia:

A la Laura, perquè més enllà d'ajudar-me a introduir curosament les referències bibliogràfiques, ha patit i gaudit d'aquest procés donant-me suport i cuidant-me com només ella sap cuidar, gràcies per ser el meu equip.

A la Begoña, que ha sabut dirigir la meva dispersió, ha tingut la paciència que a mi molts cops m'ha faltat i ha fet possible, en molts sentits, que pugui estar escrivint aquestes paraules...

Als meus pares, Blanca i Xavier, que tant han fet i que seria ridícul intentar reduir-ho a un conjunt de paraules, per la companyia i les llargues, llarguíssimes, sobretaules de converses.

A les meves germanes, Júlia, Nina, Laura i Bruna, que durant tota la vida m'han donat un recolzament incondicional, no podria ser més feliç que tenir-vos.

Als meus amics i amigues, que tanta sort tinc de poder gaudir, que m'han ajudat i animat tant, han fet que cada moment valgués la pena.

A la resta de família, que són l'alegria, la casa i el suport de la meva vida.

A la Fundació Grífols Víctor i Lucas que m'ha permès compaginar aquesta recerca amb la feina, i a l'equip que la fa possible, Maria, Núria i Sílvia, que tant m'alegro de tenir de companyes.

A les companyes i companys de l'àmbit acadèmic, que han estat un estímul intel·lectual i social al qual recórrer en escassetat d'idees i aixopluc moral.

A totes les persones que mai han deixat passar l'oportunitat de preguntar-me com portava la tesi.

No voldria acabar sense tenir un record cap a les persones que ens han deixat, amb les quals no podré compartir aquest moment i que se'm trenca el cor per no poder-ho fer: seguim junts.

Resum

La intel·ligència artificial (IA) ha penetrat profundament en diverses esferes de la vida, influenciant la manera com prenem decisions i ens comportem. Tanmateix, el seu funcionament intern, sovint opac, planteja dubtes sobre el grau de control i comprensió que tenim sobre aquestes tecnologies. Volem explorar els fonaments ontològics, epistemològics i ètics de la intel·ligència artificial (IA) a través de tres preguntes que es consideren fonamentals: què és pròpiament la IA?, quin és l'estatus epistemològic de la IA?, i què és l'ètica de la IA?

La tesi busca aclarir si les capacitats de la IA poden ser catalogades com un nou paradigma ontològic o si, simplement, representen un buit epistèmic pel que fa a la comprensió dels seus processos. Es volen tractar qüestions sobre com aquestes tecnologies capten la realitat i en quina mesura aquesta forma de captar-la guarda una relació amb el coneixement. Amb aquest propòsit, és rellevant la pregunta sobre què són els problemes i quina pot ser la contribució de la IA, perquè tota aplicació tecnològica s'insereix per a donar una solució als problemes de les persones.

En el marc del creixent desenvolupament tecnocientífic, ressona la idea que els sistemes d'IA poden pensar, i que la consciència pot ser reproduïble computacionalment. Es vol examinar si, malgrat les sofisticades respostes que generen els sistemes d'IA, aquests poden ser considerats conscients, o si simplement donen respostes basades en patrons estadístics. Amb aquest propòsit es revisa l'enfocament funcionalista per clarificar el valor de les seves contribucions i, en definitiva, examinar si té sentit parlar de consciència en entitats no biològiques.

Finalment, s'aborden les qüestions ètiques que planteja la IA, per a veure en quina mesura podem parlar de l'ètica de la IA, si és possible integrar quelcom com l'ètica en una màquina i, si fos el cas, quin tipus d'usos tecnològics haurien d'impedir-se. En darrer terme, davant la complexitat de reduir determinades activitats humanes a algoritmes, es tracta d'establir criteris per a esbrinar quines tasques poden ser delegades a sistemes d'IA i quines no haurien de ser-ho. Aquesta recerca ofereix una visió crítica de la IA, subratllant la necessitat de comprendre i regular la seva influència en la societat.

Paraules clau: intel·ligència artificial, ontologia, epistemologia, ètica, funcionalisme.

Abstract

Artificial Intelligence (AI) has deeply penetrated across a wide range of domains in our daily lives, influencing the way we make decisions and our behavior. Likewise, its often opaque inner workings, raises questions on the level of control and understanding we have over these technologies. This study aims to explore the ontological, epistemological and ethical foundations of AI through what we perceive to be three essential questions: what exactly is AI? what is the epistemological status of AI? and what are the ethics of AI?

The thesis seeks to clarify whether AI's capabilities can be categorized as a new ontological paradigm or whether it simply represents an epistemic gap in terms of the understanding of its processes. The aim is to address the issue of how these technologies capture reality and to what extent this process of capturing itself relates to knowledge. To that end, it is interesting to question both the problems that AI raises and the contributions that it can offer, since afterall, all technologies are born and introduced to bring solutions to humanity's daily life challenges.

In the context of a growing techno-scientific development, the idea that AI systems can think, and that conscience is computationally reproducible, is a common perception. This work seeks to examine whether, despite the sophisticated responses that AI systems generate, these can be considered to be conscious, or whether these are simply responses based on statistical patterns. Through this process, this thesis reviews the functionalist approach to clarify the value of AI's contributions, and ultimately to examine if it is appropriate to apply the term conscience to non-biological entities.

Lastly, this research approaches the ethical questions that AI poses, to better understand the degree to which we can talk about AI ethics, whether it is possible to integrate something such as ethics into a machine and, in such case, what uses of AI should be forbidden. Amid the complexity of reducing certain human activities to algorithms, the aim is to establish the criteria to determine which tasks should be delegated to AI and which ones should not. This research offers a critical view of AI, underscoring the need to comprehend and regulate its influence on society.

Keywords: artificial intelligence, ontology, epistemology, ethics, functionalism.

Índex

INTRODUCCIÓ 1
CAPÍTOL 1. UNA INTRODUCCIÓ A LES PROBLEMÀTIQUES DE LA IA 5
1.1. De què parlem quan parlem d'IA? 5
1.2. Definint l'objecte d'estudi 7
CAPÍTOL 2. Una ontologia de la IA 11
2.1 Una definició 11
2.2 Una categorització 14
2.3 Algunes aplicacions i regulacions 21
2.4 Conclusions sobre l'ontologia de la IA 27
CAPÍTOL 3. QUIN ÉS L'ESTATUS EPISTÈMIC DE LA IA? 29
3.1 Les persones tenen problemes, la IA també? 29
3.2 Què fa pròpiament la IA? 32
3.3 Pot dubtar la IA? Una reflexió des de Descartes 37
3.4 Pot percebre la IA? 45
3.5 El món de la vida, l'experiència i el cos 53
3.6 Conclusions sobre l'estatus epistèmic de la IA 59
CAPÍTOL 4: QÜESTIONS SOBRE EL PROBLEMA MENT-COS 61
4.1 Sobre objectivitat i subjectivitat 61
4.2 Les respostes monistes 69
4.3 Materialisme i neurociència 77
4.4 El model computacional de la ment: el funcionalisme 80
4.5 Consciència artificial i funcionalisme computacional 84
4.6 Contra el funcionalisme 92
4.7 Conclusions sobre el problema ment-cos 99
CAPÍTOL 5. EL MITE DE LA IA 101
5.1 Revisant el Test de Turing: confusions i error 101
5.2 La IA no és intel·ligència artificial 107
5.3 Conclusions sobre el mite de la IA 113
CAPÍTOL 6. ÈTICA EN LA IA 115
6.1 A què ens referim quan parlem d'ètica de la IA? 115
6.2 Qüestions ètiques des de l'ontologia i l'epistemologia 118
6.3 Pot ser ètica la IA?: límits i responsabilitat 119
6.4 Agència ètica i algoritmes 121

6.5 Contra les caixes negres 127
6.6 Conclusions sobre l'ètica en la IA 130
CONCLUSIONS 133
REFERÈNCIES BIBLIOGRÀFIQUES 137

INTRODUCCIÓ

Lluny d'haver estat anunciada, la irrupció de la intel·ligència artificial (IA) s'ha precipitat ràpidament sobre totes les esferes de la vida. Els algoritmes de la IA s'han convertit en els nous gurus als quals tan sovint es recorre per orientar la conducta humana. Sota el rètol d'intel·ligents, els nous sistemes tecnològics es presenten com l'obsolescència d'aquelles capacitats que creiem identitàries de l'espècie humana: l'intel·lecte, la raó, la consciència i el pensament. Fins i tot, es parla de sistemes autònoms que poden adaptar el seu comportament a diversos contextos. La llibertat també sembla caure sota el domini de les màquines, les quals programem perquè puguin anticipar els nostres interessos i satisfer, així, els nostres desitjos (també suggestionar-nos). No ens sembla estrany delegar la presa de decisions en multitud de tasques de la nostra vida a algoritmes, també aquelles decisions que tenen a veure amb la reflexió ètica. Anàlogament al que Walter Benjamin havia dit sobre l'obra d'art, ara podem parlar de l'ètica a l'època de la seva reproductibilitat tècnica

Ha canviat la tecnologia i aquesta ha transformat la vida humana. Tanmateix, el funcionament dels artefactes equipats amb IA no és diferent d'altres màquines. Si abans es basaven en processos mecànics, ara ho fan en el processament de grans quantitats de dades (*big data*) i l'extracció de models estadístics. L'autèntica transformació és que ara podem aplicar aquests models estadístics perquè prenguin les nostres decisions i resolguin els nostres problemes. La creixent sofisticació dels sistemes d'IA permet que generin respostes efectives o útils per a la tasca en qüestió. No obstant això, sovint desconeixem els processos interns que han seguit per arribar a aquestes respostes, donat que poden funcionar com a caixes negres que dificulten entendre i rastrejar amb precisió el raonament o les dades que han utilitzat.

En aquesta tesis volem argumentar si ens trobem en un nou paradigma ontològic de les màquines, o es tracta més bé d'una escletxa epistèmica sobre el seu funcionament. Hem d'examinar si la tecnologia està en condicions de prendre les nostres decisions, i identificar quin tipus de decisions és inacceptable delegar perquè resulta especialment complicat reduir-les a un algoritme. Això ens permetrà avaluar les limitacions de la IA i establir quin pot ser el paper de la IA en la resolució dels nostre problemes.

L'objectiu que persegueix aquesta tesi és respondre i plantejar qüestions fonamentals sobre les noves tecnologies basades en IA. El pressupòsit que es vol demostrar és que si bé la IA pot imitar el comportament humà, no operar d'acord amb aquelles activitats que identifiquen el que és l'ésser humà. En aquest sentit, ens centrarem en l'anàlisi de la IA des de l'ontologia,

l'epistemologia i l'ètica. En concret volem respondre a les següents preguntes: què és pròpiament la IA?, quin és el seu estatus epistemològic? i què és l'ètica de la IA? Volem analitzar els aspectes ètics de la IA, la qual cosa ens porta abans a conèixer quines són aquelles tasques que la IA pot fer i com les porta a terme; i si volem conèixer el que pot fer, abans hem de definir què és i quin tipus d'entitat pot ser catalogada com a IA.

Aquest treball consta de 6 capítols que s'estructuren de la següent manera:

En el Capítol 1, per fixar l'objecte d'estudi d'aquesta tesi, es fa una introducció sobre quina és la problemàtica al voltant de la IA i de què parlem quan parlem d'IA.

En el Capítol 2, es vol clarificar la identificació i categorització del què és la IA, aportant una definició que ajudi a reflexionar i a entendre-la millor. La importància de proporcionar aquesta definició és que la imatge que es té de la IA sol desviar-se del que realment és. Aquesta és una conseqüència dels relats que ens ha presentat la ciència-ficció, o veus entusiasmades amb el desenvolupament tecnològic (tecnofílics) o, al contrari, preocupades pel domini de les màquines intel·ligents (tecnofòbics). Donat que els sistemes d'IA han estat posats al món de forma precipitada i massiva i és necessària una regulació, presentarem algunes de les aplicacions de la IA i la seva regulació jurídica.

En el Capítol 3, com que tota aplicació tecnològica s'insereix per a donar una solució als problemes de les persones, ens dediquem a la pregunta sobre què són els problemes i quina pot ser la contribució de la IA a aquests. Amb aqueta propòsit i arran de les múltiples aplicacions que es dona als sistemes d'IA, analitzem com aquestes tecnologies capten la realitat i en quina mesura aquesta forma de captar-la guarda una relació amb el coneixement humà. A partir d'aquesta qüestió es vol indagar en què és allò que pròpiament pot fer una IA. Posant el focus d'atenció en activitats humanes com la de jugar, veurem si la gran competència que demostren alguns programes informàtics per dominar alguns jocs (tals com els escacs o el Go) significa comprensió. Per aprofundir en el tema, es proposa una aproximació des del mètode cartesià del dubte i la percepció per a respondre a si la IA pot tenir percepcions quan modelitzen el seu entorn, les tasques que han de realitzar i de quina forma resolen les tasques encomanades.

En el Capítol 4, abordem un dels problemes clàssics de la filosofia: el problema ment-cos. L'objectiu és fer aportacions al controvertit debat sobre si es pot atorgar propietats mentals a les tecnologies basades en IA. En el context del creixent desenvolupament tecnològic, es parla de sistemes d'IA conscients i que considerem com entitats pensants. Alimentades per les troballes

en neurociència, algunes opinions sintetitzen tot el que és l'ésser humà a un cervell, creient que ha de ser reproduïble computacionalment i que en fer-ho es disposarà d'una copia digital (no biològica) de la consciència. Per això revisem l'enfocament funcionalista per clarificar el valor de les seves contribucions i, en definitiva, examinar si té sentit parlar de consciència per a entitats no biològiques.

En el Capítol 5, partint de les tesis funcionalistes i conductistes, hi ha un seguit de prejudicis al voltant de la IA que hem identificat sota el lema el mite de la IA, que han de ser estudiats a la llum de les reflexions que es desprenen dels capítols precedents. Un d'aquests mites és que el Test de Turing ofereix una definició d'intel·ligència segons la qual ser intel·ligent és comportar-se intel·ligentment. Un altre mite és que la IA és intel·ligent, un aspecte que es vol analitzar a través d'arguments que contribueixin a una concepció de la IA menys estereotipada i antropomorfitzada.

En el Capítol 6, s'atenen les qüestions ètiques sobre la IA. A partir de les reflexions ontològiques i epistemològiques exposades en els capítols anteriors, l'objectiu en aquest cas és orientar i contribuir al debat al voltant de què és l'ètica de la IA, en quin sentit l'ètica es pot integrar en una tecnologia, si podem parlar d'agència ètica d'una màquina i quins usos tecnològics hauríem d'evitar.

La tesi no se centra en una perspectiva concreta sobre la IA, ja que es tracta d'una temàtica que no està tancada i que es presenta des de diversos enfocaments. El que sí que es vol fer és assenyalar aquelles concepcions de la IA que en la seva explicació poden resultar desencaminades al que són i poden fer els sistemes d'IA.

CAPÍTOL 1. UNA INTRODUCCIÓ A LES PROBLEMÀTIQUES DE LA IA

L'ésser humà es caracteritza —i podem dir que es distingeix de la resta d'animals— per orientar la seva vida segons la imatge de qui o què és, de qui o què ha de ser, és a dir, que disposa d'una imatge d'ell mateix. Som els únics que podem donar resposta a les preguntes què o qui som i què o qui hem de ser. Dit d'una altra manera, l'única manera de conèixer l'ésser humà és essent humà.

El camp de la problemàtica al voltant de la IA té implicacions en àmbits centrals de la societat, un canvi radical que ha vingut fortament marcat per la revolució tecnològica dels darrers anys. Aquesta també transforma la manera de relacionar-nos entre nosaltres —incloent la resta d'animals i ecosistemes— i de relacionar-nos amb la tecnologia. Un fet que té conseqüències a l'hora d'avaluar els riscs de l'era digital on el debat sobre les tecnologies basades en IA es mou entre dos extrems. Des dels aterradors escenaris d'una superintel·ligència artificial amenaçadora, fins als transhumanistes que veuen en la tecnologia un mitjà per transcendir la nostra biologia. Ens trobem davant d'un procés que està truncant la imatge que tenim de nosaltres mateixos i la de com hauríem de ser.

1.1. De què parlem quan parlem d'IA?

El pensament, la ment, és la nostra capacitat insígnia, la que ens distingeix de la resta d'animals i objectes, la que ens permet crear una imatge de nosaltres mateixos i és a l'arrel del que anomenem intel·ligència. En el cas que aquestes afirmacions siguin vertaderes, de quina manera es podria arribar a afirmar que existeix *intel·ligència artificial*, és a dir, una tecnologia capaç de sentir, pensar i viure —qui sap si— d'una forma diferent a la nostra? Ens veuríem obligats a afirmar que també disposa de la capacitat de pensar, i per tant, de crear una imatge de si mateixa, de qui i què és i ha de ser.

La informàtica ha heretat la forma de pensar de la lògica i les matemàtiques —un mèrit que segurament devem a Frege, Russell i Wittgenstein—, i la forma de pensar de la lògica i les matemàtiques elabora un model del pensament humà, és a dir, una estructura suficientment isomorfa amb l'estructura del pensament humà, però sense arribar a copiar-la. La lògica no és una fotografia dels processos mentals, és un model de com s'ha de pensar. La lògica construeix un ideal de la racionalitat, que, tanmateix, no és una descripció o una explicació dels processos del que és ser humà, sinó que consta del seguit de normes per a l'elaboració d'una forma de pensar. D'aquí es deriva que mai no hi podrà haver una IA que sigui indistingible del pensar humà sota tots els seus aspectes, perquè el mapa mai serà igual al territori. Ara bé, el que sí que pot

és modificar el pensar humà descobrint-ne noves propietats (topogràfiques), ja que s'estableix una relació de retroalimentació entre el territori i el model que condiciona la nostra realitat. I és en aquesta relació en la qual rau l'autèntica potencialitat, tant pel que fa els beneficis com perills, de la IA, i no en el fet que aquests models ens superin.

El territori del pensar es modifica a través de la digitalització, ja que aquest procés consumeix grans quantitats d'energia. La quantitat d'energia que consumeixen els sistemes d'IA té un gran impacte mediambiental, molts cops invisible, i també contribueix significativament a la crisi ecològica actual. Cada clic és un procés físicament mesurable i, per tant, un consum de les reserves d'energia. Cada internació en l'espai digital té una empremta energètica, per la qual cosa amb cada clic contaminem literalment el nostre medi ambient de forma imperceptible. Però no ens aturem a pensar en aquest fet perquè l'ús de la IA ens produeix la sensació d'elevar-nos per sobre la nostra existència material, fent-nos caure en una espècie de platonisme digitalitzat aliè al materialisme històric.

La IA no pensa com nosaltres, ni està en competència amb nosaltres pel senzill motiu que nosaltres no programem els nostres interessos. Si el meu mòbil intel·ligent em recomana productes que s'ajusten a la meva forma de pensar —segurament modelitzada pel mateix algoritme—, no vol dir que aquest artefacte tecnològic vulgui anar de compres. La IA, doncs, no s'orienta a si mateixa segons les recomanacions que proposa. No obstant això, precisament per aquesta estructura funcional, existeixen perills reals que són una gran amenaça. Tal amenaça ve alimentada pel fet que l'ètica no és reduïble a un algoritme, a un càlcul logicoformal. La nostra manera particular de prendre decisions ens protegeix contínuament de la fallida social, no perquè siguem totalment racionals, sinó perquè la nostra ment no és essencialment un programa predefinit.

No pensem i actuem sota condicions operatives ètiques determinades per un càlcul que deriva judicis morals com si de teoremes es tractessin. Per això tota teoria ètica general xoca amb paradoxes insuperables, cosa que té conseqüències per a la seva implementació en una IA, ja que posa al descobert que no som capaços de concebre una imatge clara i general de la nostra pràctica moral per tal de traslladar la nostra autonomia a un programa autònom.

La conceptualització dels primers sistemes computacionals *intel·ligents*, en gran part duta a terme per Alan Turing, es basa en que només és legítim suposar que algú altre és intel·ligent identificant el seu comportament observable, és a dir, observant que es comporta de forma intel·ligent (com es comporten les persones quan diem que s'estan comportant

intel·ligentment). Segons això, realment no existeixen ni la intel·ligència, la consciència o el pensament, només existeix una forma de comportar-se anàloga a posseir tals capacitats i l'observació d'aquest comportament. Aquest és el tret principal del conductisme: si no podem distingir entre la realitat de la intel·ligència, la consciència i el pensament de la seva mera adscripció, com podem saber si certs artefactes amb els quals interactuem no són intel·ligents? Estrictament parlant, no és que sigui una contradicció pensar que els sistemes d'IA puguin ser intel·ligents, simplement no tenim arguments per justificar-ho.

1.2. Definint l'objecte d'estudi

La irrupció de sistemes basats en IA, que cada cop abasten més àmbits de la vida humana, ha obert noves preguntes sobre què o qui som i ha exacerbat antigues problemàtiques que mai havien quedat resoltes. Aquestes engloben qüestions que fan referència a aspectes tecnocientífics per a integrar nous models algorítmics, passant per la forma en què la societat es comunica en l'era digital, en els usos que donem a la tecnologia, fins a arribar a les implicacions filosòfiques i ètiques que es deriven d'aquests fenòmens. En un món que canvia amb cada clic, amb cada nova cerca, amb cada segon d'atenció i interacció, pot semblar que s'està perdent la perspectiva de la realitat. Una realitat en la què cada cop ens costa més parar atenció a allò que està passant fora d'unes pantalles alimentades amb ingents quantitats de dades generades per nosaltres. Ens hem adaptat amb facilitat i rapidesa a l'entorn digital. No obstant això, l'adaptació digital no ha estat seguida per una reflexió social sobre el seu abast, i això comporta que l'esfera social vagi a remolc de la tecnològica. Cal destacar també que, lluny de sempre solucionar els nostres problemes, les noves tecnologies contribueixen a fer-los encara més grans. Els problemes mediambientals, socials i ètics no se solucionen amb algoritmes més sofisticats i computadors cada cop més potents. Amb cada clic, amb cada nova cerca, també consumim immenses quantitats de recursos energètics per alimentar l'entorn digital en el quan ens movem, encara que no ens en adonem. La digitalització no està contribuint a resoldre els nostres problemes morals i social, més aviat els ha agreujat, sovint amb la nostra (in)conscient complicitat.

Des dels anys seixanta, els humans, la nostra fisiologia i el nostre cervell, no han evolucionat gaire. En canvi, sí que ho ha fet notablement la tecnologia que hem dissenyat, o millor dit: la nostra capacitat per a fabricar-la. Més o menys, els cotxes actuals són el doble de ràpids que els de la dècada del 1960, afegint que ara ens poden guiar per arribar al nostre destí i disposen d'un grapat de prestacions més que tan sols uns anys enrere ens haguessin semblat una ficció. No és estrany que, en dubtar sobre el significat d'una paraula, puguem desembutxacar els nostres mòbils intel·ligents i resoldre el misteri amb una senzilla cerca o preguntant-li al bot

conversacional *ChatGPT*. Però, amb certesa, el que ha evolucionat més fins a dia d'avui ha estat la potència de processament de dades (s'estima que en un bilió de vegades), absolutament cap altra cosa ha millorat a aquest ritme. El 90% de totes les dades que existeixen s'han creat en els darrers cinc anys (Du Sautoy, 2020, p. 86) i a diari es produeixen 328.77 d'exabytes.[1] Per tenir una idea aproximada de l'enorme volum d'informació que això suposa, en l'actualitat en dos dies es genera la mateixa quantitat de dades que la generada des dels inicis de la civilització fins el 2003. La capacitat de processar dades ha suposat el desenvolupament dels sistemes basats en IA, que incorporen multitud de màquines que fem servir i ens assisteixen quotidianament en la gran majoria de tasques que realitzem (fer la compra *en línia*, suggerir noves cançons basant-se en els nostres gustos, la conducció de vehicles, etc).

Ha estat tal el progrés dels darrers anys que les aplicacions de la IA acaparen gran part dels àmbits on interactuem. La logística de transports, els sistemes de reconeixement facial, les finances, la publicitat, els serveis sanitaris i socials, l'educació, tots ells són sectors on es fan servir eines que aprofiten aquesta tecnologia. Si tenim en compte que el cas del reconeixement d'imatges, que es pot utilitzar per a diagnosticar precoçment malalties, semblaria difícil qüestionar els avantatges que ens aporta la IA. Malgrat això, tot es complica en el moment en què assenyalem la quantitat de decisions sobre les nostres vides que prenen aquests algoritmes que es basen en IA. Les preguntes que això ens suscita són múltiples. Podem confiar més en el procés de presa de decisió de la IA que en el nostres? Hauríem d'escollir quines decisions volem *delegar* a la IA i quines altres de cap de les maneres? Qui es responsabilitzarà dels resultats de les decisions si alguna cosa surt malament? Quin grau de risc estem disposats a assumir? Tampoc no hem de perdre de vista que una faceta de les noves aplicacions tecnològiques —alineat amb el desig de rendibilitat econòmica— és la de crear necessitats, així pot resultar molt més fàcil satisfer-les, o millor dit, pot resultar molt més fàcil saber fins a quin punt satisfer-les. És complex saber on acaba la necessitat i on comença la comoditat en aquests usos, quan ens manca l'avantatge històric per a jutjar-ho i hem d'aventurar les conseqüències que podrien tenir.

El sistemes d'IA, apunta Markus Gabriel (2019, p. 130), són una amenaça per a la humanitat, perquè ens recomanen implícitament els sistemes de valors de les persones que els han programat o finançat, sense fer-los transparents; és a dir, sense que tinguem accés a les seves motivacions morals. Així, el disseny dels algoritmes basats en IA persegueix una moral determinada, una visió de com ens hauríem de comportar, alhora que se'ns presenten com un

[1] Cada exabyte equival a 10^{18} bytes de dades.

càlcul de valors neutre, al qual es pot tenir accés de forma senzilla. Tinguem present, tanmateix, que no hi ha valors neutres, ni tampoc es pot ser èticament neutre.[2] No podem desfer-nos dels nostres perjudicis i biaixos perquè en cada acció despleguem valors morals (P.-Paul. Verbeek, 2011), és el que Max Weber defineix com a *neutralitat axiològica*.[3]

El disseny dels sistemes d'IA que fem servir contenen biaixos que condicionen —moltes vegades deliberadament— la nostra conducta, un fet del qual alerta la matemàtica experta en anàlisi i gestió d'informació, Cathy O'Neil (2018), «els algoritmes són opinions posades en codi i no són objectius». A través de l'aparent neutralitat dels algoritmes, som guiats en les nostres incursions per l'entorn digital. Però també som guiats pels objectius, ideologies i prejudicis en els quals es troben confinats els programadors i empreses que financen aquests algoritmes. A grans trets, són molts els interrogants que ens conviden a examinar detingudament les implicacions d'aquesta tecnologia, i en especial, pel seguit de qüestions referents a la responsabilitat, la llibertat, l'agència moral i la dignitat que plantegen. Encara que les motivacions per a desenvolupar noves tecnologies siguin les més filantròpiques, com a resultat de la nostra interacció amb elles sorgeixen conseqüències indesitjables i sovint inesperades, donant pas a nous problemes ètics o reobrint-ne de passats. És urgent educar a les generacions actuals i futures sobre els usos tecnològics responsables, perquè, tot i no afrontar-ne els mals usos, en patirem les conseqüències (Latorre Sentís, 2019, p. 215).

Hem d'assegurar-nos que el progrés tecnològic sigui una eina democràtica en beneficis de tota la societat, en comptes de contribuir a fer més grans les desigualtats socioeconòmiques i només estar al servei de les grans empreses que les desenvolupen (Nemitz, 2018). El terreny en el qual s'ha de lliurar aquesta discussió ha de ser el de l'ètica, ja que aquests debats són encapçalats per preguntes tals com «per què hauríem —o no— d'actuar d'aquesta forma?», una pregunta essencialment ètica. És precisament aquesta actitud crítico-reflexiva de l'ètica la que ha d'orientar el desenvolupament tecnològic i als seus usuaris. Ara bé, la potència disruptiva de la IA ens situa en la tessitura de fer prèviament un examen filosòfic exhaustiu d'aquesta tecnologia

[2] Es podria argumentar, tal com fa Markus Gabriel, que hi ha accions que no es poden catalogar com a moralment correctes o incorrectes: muntar amb bicicleta, torrar el pa, caminar per la vorera o fer un glop d'aigua.

[3] Seria convenient que en fer estudis sobre IA es distingissin les nocions acció i comportament, entenent-les com les capacitats arendtianes d'obrar (actuar) i de fer (comportar-se), com proposa la companya Júlia Pareto (2021). En aquest treball s'assumeix que la capacitat d'actuar és exclusivament humana, ja que està fonamentada en l'autonomia moral i la intencionalitat, mentre que la conducta s'associa amb l'autonomia fàctica o capacitat d'execució, absent d'intencionalitat, que és el que els artefactes tecnològics poden fer en el millor del casos.

per poder entrar a discutir-ne els aspectes ètics més rellevants. Si volem fer ètica de la IA, primer hem de definir el nostre objecte d'estudi. No podrem entrar a valorar quins poden ser els bons i mals usos d'una eina sense abans establir de què estem parlant. És des de l'ontologia que podem saber pròpiament parlant què és la IA, i és des de l'epistemologia que la podrem entendre.

Els professionals del sector tecnològic sovint utilitzen nocions filosòfiques per a explicar el funcionament de les seves invencions, excedint així el seu camp d'expertesa —limitat a un caràcter més tècnic i logicoformal— per falta d'eines conceptuals. Generen així confusions al voltant dels conceptes filosòfics als quals recorren per a explicar el funcionament de les màquines que dissenyen. Els enginyers, programadors, científics de dades i especialistes en ciberseguretat s'escapen de l'àmbit estrictament científic i, segurament sense saber-ho, es converteixen en filòsofs. Mentre que termes tals com *software*, *hardware* i *robot* formen part del marc conceptual tecnocientífics, els d'*intel·ligència* i *artificial* també formen part del llenguatge filosòfic. Per això, les idees filosòfiques es poden veure separades del seu significat, en mans de tecnòlegs que —acostumats a fer servir conceptes rígids— traslladen descripcions purament logicoformals a contextos no tècnics que escapen del seu domini com la psicologia, la biologia o la neurologia. Aquest escenari ens motiva a buscar una fonamentació filosòfica sobre els conceptes que es fan servir en l'àmbit de la IA, per tal d'identificar els errors, corregir les confusions i clarificar els conceptes que s'hi veuen involucrats.

Comptat i debatut, la nostra pregunta és què és això que avui s'anomena intel·ligència artificial? Per fer una ontologia de la IA és important tenir una definició i una organització categorial dels elements que la conformen, però també és central examinar les seves aplicacions pràctiques i com seran regulades. A això ens dedicarem a continuació.

CAPÍTOL 2. Una ontologia de la IA

Identificar l'ontologia de la IA vol dir distingir el conjunt de termes formals que la conceptualitzen i que en representen el coneixement (Gruber, 1993). A què ens referim quan parlem d'IA? En què consisteix? Quines categories se'n poden distingir? Què se'n sap? Ens proposarem aquí oferir una definició sobre com pot ser entesa la IA, assumint que no podem reduir tots els conceptes que acull a una única definició. És per això que optem per parlar d'*una* definició —fent servir un pronom indeterminat—, perquè encara que estem parlant d'una cosa concreta que volem definir, hem d'entendre que no és l'*única* forma de fer-ho. Es tracta, doncs, de suggerir una noció de la IA que ajudi a entendre i a reflexionar millor sobre aquesta disciplina.

2.1 Una definició

Diversos estudiosos de la IA (Coeckelbergh, 2021, p. 62; Kaplan, 2017, p. 8-9; McCarthy et al., 2006; S. Russell & Norving, 2010) coincideixen en que pot ser definida com una disciplina científica —com la física o la geologia— i, per tant, com una col·lecció de conceptes i problemes amb un seguit de mètodes per a resoldre'ls. Entesa així, la IA forma part del camp de la informàtica destinat al desenvolupament d'algoritmes que tenen com a objectiu dur a terme tasques concretes que normalment requeririen de la intel·ligència humana per a ser realitzades.

Tot i que la descrita només és una de les facetes de la IA, ens situa en el moll de la qüestió, ja que trasllada a aquests sistemes autònoms la nostra autonomia fàctica[4] però no l'autonomia moral, i cal distingir-les (Etxeberria i Casado, 2008). L'autonomia moral fa referència a la capacitat racional suficient per a prendre decisions (actuar), que respon a una voluntat pròpia i al fet de poder respondre de les conseqüències de les accions (responsabilitat moral). En canvi, l'autonomia fàctica o tècnica (capacitat d'execució) consisteix a dur a terme determinades conductes. No diferenciar correctament ambdós conceptes porta a confusions i la importància de distingir-los rau en evitar situar els sistemes d'IA entre les entitats ontològicament individuals (Coeckelbergh, 2011). Aquestes imprecisions i els equívocs als quals deriven, entorpeixen la reflexió ètica perquè conceben un objecte —una eina en aquest cas— amb un subjecte tot atorgant a una màquina l'estatus d'agent moral que pot danyar i ser danyat.

Els responsables de què els nostres ordinadors, mòbils i sistemes GPS funcionin són els algoritmes. Quan fem una compra *en línia* i dies després ens apareixen anuncis i suggeriments

[4] En la literatura sobre IA i robòtica s'acostuma a utilitzar autonomia tècnica quan es parla d'autonomia fàctica, per al·ludir al que poden fer les eines tecnològiques i distingir-la de l'autonomia humana.

—als webs que visitem, a les xarxes socials, etc.— sobre altres productes que estan relacionats amb la compra que vàrem fer, és gràcies a un algoritme que ha estat dissenyat per modelitzar el nostre comportament a fi de —en el millor dels casos— preveure quin producte ens podria interessar, o —en el pitjor— suggestionar-nos per a comprar-lo. Però els algoritmes no són res nou i van aparèixer molt abans que els ordinadors existissin.[5] Dit de manera general, un algoritme és una regla que estableix que els passos ben definits d'un procés per obtenir un resultat o una solució a un problema donat. Així mateix, hem de tenir present que per tal que un algoritme funcioni no és necessari que estigui basat en cap mena d'IA.

Els algoritmes es defineixen per les seves propietats lògiques i cal aturar-nos uns instants en aquest fet per destacar un dels grans mèrits d'Aristòtil i els seus contemporanis. El valor de les seves aportacions va ser concebre un model de realitat i de pensament que s'encarrega d'estudiar la forma amb què opera el raonament —a través de regles— per a determinar si un argument és vàlid o no. La lògica descriu les condicions sota les quals els continguts del pensament es connecten entre si i amb la realitat. Des d'un punt de vista lògic, té la mateixa rellevància el contingut d'un pensament que anuncia «Plató va ensenyar a Aristòtil» que «Aristòtil va ser ensenyat per Plató», són continguts lògicament idèntics. D'aquí no només es dona a conèixer la relació entre els dos pensadors, sinó que també es dedueix que existeix una persona que va ensenyar a Aristòtil i una persona que va ser deixeble de Plató. I, així mateix, les lleis de la lògica també descriuen les condicions de control dels algoritmes. La lògica ens proporciona una mena de manual per a traduir la realitat —conceptualitzant-la— a un llenguatge que ens sigui comprensible. Ara bé, la lògica tan sols afirma quelcom sobre la realitat, sense proporcionar-nos una descripció del que és en conjunt. En altres paraules, la lògica estableix les bases per a la digitalització —parcial— de la realitat i és cabdal per què els programadors puguin dissenyar els algoritmes basats en IA.

Certament, sembla complicat distingir on acaba un algoritme basat en IA i on comença la ment del programador, amb les seves intencions, prejudicis i objectius. El problema és que ja d'entrada és difícil distingir el que és un sistema d'IA del que no i això pot resultar complicat per diversos motius. En primer lloc, perquè generalment ens estem referint a tasques molt diverses que poden realitzar els sistemes d'IA, i malgrat que s'hagin establert certes funcions d'aquests, no hi ha un acord oficial sobre la seva definició. En segon lloc, gran part de les visions estereotipades

[5] Els primers algoritmes dels quals es té constància es remunten al primer terç del segon mil·lenni abans de Crist (entre el 2000-1600 aC), i els feien servir els babilonis per a calcular molt eficaç i ràpidament arrels quadrades a mà (Fowler & Robson, 1998).

que es tenen de la IA són l'herència d'una ciència-ficció que ens ha presentat robots humanoides (*Metrópolis*), l'alçament dels robots contra els humans (*Jo, robot*), un sistema d'IA autoconscient (*2001: Una odisea espacial*), replicants capaços de desenvolupar sentiments (*Blade Runner*), una xarxa de computació global (N*euromante*) i tota mena de màquines que, per poc *intel·ligents* que siguin, amenacen la continuïtat de la vida humana a la terra. En tercer lloc, hi ha accions que per les persones són molt senzilles de realitzar (desplaçar-se per un entorn canviant) i, per contra, són extremadament complicades de programar per tal que les dugui a terme una màquina. I, finalment, mentre que hi ha accions que per a les persones requereixen un gran esforç, concentració i temps (operar amb grans quantitats de dades), fa molt temps que existeixen màquines que superen a les persones en capacitat i velocitat de càlcul, processament de dades o reconeixement de patrons.

Quedem fascinats per l'enorme eficiència computacional que tenen els nostres ordinadors i la rapidesa amb què resolen operacions que a nosaltres ens portarien molt més temps. Però, al cap i a la fi, no hi ha gaire misteri en resoldre aquets tipus de problemes, i ser capaç de resoldre un problema donat no ens diu res sobre l'habilitat per a resoldre'n un de diferent. Tot i que no hi hagi un consens generalitzat, alguns experts convenen en definir el sistemes d'IA com a algoritmes que aconsegueixen resoldre un problema d'una manera original i que els humans som incapaços de comprendre . D'acord amb aquesta concepció, no ens podem referir a IA quan parlem d'algoritmes que segueixen una seqüència predeterminada d'ordres en les quals hem bolcat tota la informació de què disposem sobre la qüestió que volem que analitzi o la tasca que volem que resolgui.[6]

Des del punt de vista de l'enteniment humà, no és cap enigma el processament veloç i eficient de grans quantitats d'informació. Podem entendre perfectament per què tarda tan poc en fer colossals successions de càlculs, com funciona internament i saber el temps que tardarà en fer-les. Sí que ens sorprendria o espantaria, en canvi, ordenar a un programa que fes un càlcul i observar com el resol molt abans que haguéssim acabat de donar l'ordre, o que, fins i tot, fos capaç d'anticipar les properes ordres que li anéssim a donar. Tot això, sense aconseguir entendre de quina manera funciona per poder anticipar-se el nostre comportament. Fora d'aquesta escenari improbable, sí que existeixen actualment sistemes d'IA tan opacs que dificulten molt poder conèixer els mecanismes que porten a terme internament.

[6] Malauradament, aquesta definició també condueix a acceptar com a IA únicament aquells sistemes que funcionen sense que puguem conèixer com ho fan, és a dir, les caixes negres, que veurem tot seguit.

Les *caixes negres* (*Black boxs* en anglès) és el nom que reben els algoritmes d'IA dels quals només sabem quina entrada ha rebut i quina sortida ha donat, però no de quina forma arriba a una conclusió o en base a què pren una decisió i no una altra. Per posar-ne un exemple, són algoritmes que poden fer prediccions altament fiables, que es compleixen un tan per cent elevat de les vegades, però que som incapaços de conèixer com ho fan. Encara que generalment són molt eficients, tenen la particularitat de fer molt complicat desxifrar per quin motiu donen una resposta o una altra davant d'un *input*, però no impossible, ja que examinant les complexes arquitectures internes que tenen i aplicant diferents mètodes (segons el tipus d'algoritme) és pot conèixer quin és el mecanisme a través del qual ha operat. No tots els algoritmes basats en IA són caixes negres; que ho siguin o no, depèn de la traçabilitat del seu funcionament intern i això dependrà de la seva complexitat i disseny algorítmic (Doshi-Velez i Kim, 2017; Rudin, 2019).[7] Que l'algoritme tingui un disseny o una arquitectura determinada dependrà de l'ús que se li vulgui donar.

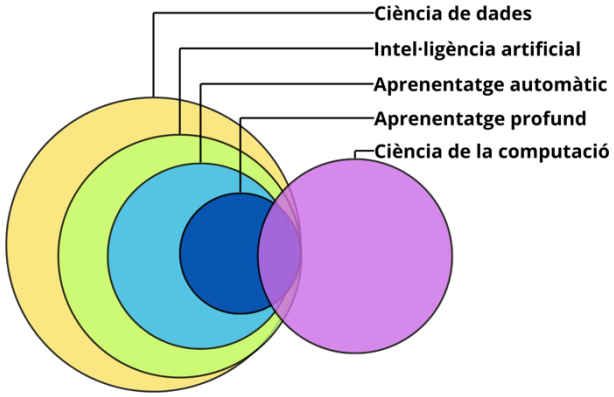
La IA es pot identificar amb un conjunt de mètodes per a processar dades a través d'algoritmes o regles, tot i que hi ha dissemblances al voltant del que poden fer. Alguns experts apunten a que per ser considerat IA, un sistema ha d'oferir una forma nova de resoldre un problema (Legg i Hutter, 2007), altres consideren que ha d'assimilar-se al comportament que tindria un humà (Turing, 1950) i altres que ha de ser capaç de fer tasques d'igual forma a com les faria una persona (Newell i Simon, 1976). El que sí que acumula més consens és que el funcionament de certs algoritmes al processar dades pot arribar a ser tan complex que requereixi un examen exhaustiu per poder-lo entendre —que no vol dir que sigui impossible d'entendre—. Els sistemes d'IA es poden classificar segons la forma com funcionen i s'estableixen les regles que els governen, però la idea general és que són programes amb la capacitat d'aprendre i adaptar-se, capacitats vinculades amb l'autonomia, com veurem. Tals capacitats els fan especialment útils per a la recollida i anàlisi de dades en contextos on aquestes dades són canviants, complexes o abundant. La categorització de les diverses branques de la IA i de quina forma estan relacionades, es basa en la forma com processen les dades i el problema que volen tractar.

2.2 Una categorització

Si l'entenem com a disciplina científica, la IA està relacionada i forma part de dos camps principals, la *ciència de dades* (*Data science*) i la *ciència de la computació* (*Computer science*), i dona lloca un subcamp conegut com a *aprenentatge automàtic* (*Machine learning*) que, al seu

[7] Un fet que contradiu la definició que proposa Latorre.

torn, conté el subcamp de l'*aprenentatge profund* (*Deep learining*).[8] La ciència de la computació se centra en estudiar teòricament els algoritmes, les dades i la complexitat computacional. Per altra banda, la ciència de dades és un camp interdisciplinari que utilitza diverses tècniques i processos algorítmics per extreure informació a partir de les dades. Com a subcamp de la IA, l'aprenentatge automàtic es basa en la idea que els sistemes poden aprendre a partir de les dades, establir correlacions (patrons) i prendre decisions, tot això, sense pràcticament intervenció humana. I, per últim, l'aprenentatge profund forma part de l'aprenentatge automàtic i utilitza *xarxes neuronals artificials* amb capes profundes de neurones per tal de modelitzar patrons complexos en les dades. Des del punt de vista de resultats, l'aprenentatge automàtic ha estat la part més fructífera a l'hora de dissenyar algoritmes i, dins aquest, l'aprenentatge profund és la tecnologia basada en IA més eficient que trobem a dia d'avui (veure Il·lustració 1: diagrama relació IA).



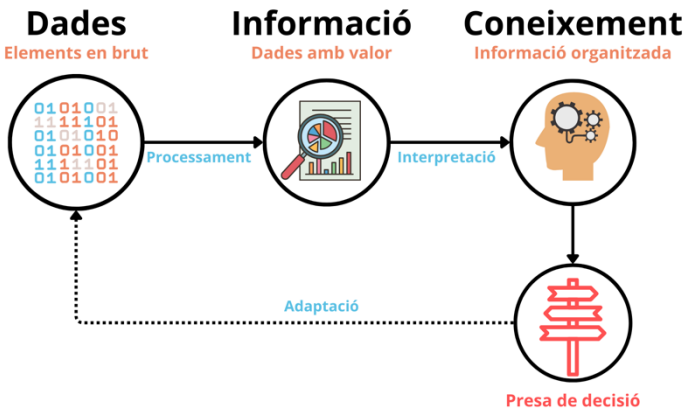
Il·lustració 1: diagrama relació IA

Font: elaboració pròpia

Als sistemes que funcionen a través de l'aprenentatge automàtic se'ls acostuma atribuir la capacitat d'aprendre, una concepció que no troba consens en els experts del sector tecnològic, ja que hi ha crítics que argumenten que no es pot parlar d'aprenentatge genuí. El motiu és que per parlar d'aprenentatge hi hauria d'haver cognició real, i aquest tipus de sistemes guarden molt poca semblança —o cap— amb el que experimenten els humans quan aprenen (M. A. Boden, 2016, p. 46). Sigui com sigui, els algoritmes basats en aprenentatge automàtic són molt útils per a establir correlacions entre dades, donat que basen el seu funcionament en l'estadística, la ciència d'extreure *coneixement* de les dades. El seu funcionament permet que un

[8] Hi ha un sector crític amb aquest punt de vista i que sosté que no tot l'aprenentatge automàtic ha de ser considerat part de la IA, només les aplicacions que fan servir tecnologies intel·ligents formen part d'un subconjunt de la IA. Per més informació sobre aquesta discussió veure (Tiwari et al., 2018).

sistema pugui identificar fàcilment patrons *observant* conjunts de dades (elements en brut), que després processen per a poder extreure'n informació útil (dades amb valor com prediccions, recomanacions o classificacions, per exemple); aquesta informació és organitzada mitjançant la interpretació, i s'obté un coneixement que guia la presa de decisions informades. Alhora, hi ha un procés de retroalimentació que possibilita l'adaptació de la recopilació i processament de les dades futures (veure Il·lustració 2: diagrama relació dades, informació i coneixement). Degut a aquest comportament, es diu que l'algoritme aprèn. El cert és que la intervenció humana és present en tot el procés d'*aprenentatge*: definir les dades que s'han de recopilar, dissenyar com es farà el processament de les dades i aplicar la interpretació de la informació a un context.



Il·lustració 2: diagrama relació dades, informació i coneixement

Font: elaboració pròpia

Segons l'orientació que necessiten en la recollida de dades, es poden desglossar el sistemes d'aprenentatge automàtic. En l'*aprenentatge supervisat*, els algoritmes necessiten que se'ls presenti un gran volum de dades i s'etiqueti la resposta correcte (quines imatges són un gat, per exemple) en una fase que es diu entrenament, amb l'objectiu de donar una sortida consistent (identificar on hi ha un gat i on no), establint patrons entre les dades d'entrenament i les noves —o futures, en el cas de la predicció— dades. L'*aprenentatge no supervisat* no necessita que s'etiquetin les dades prèviament, de manera que l'objectiu de l'algoritme no està definit, més enllà de trobar conjunts o singularitats entre les dades. En l'*aprenentatge per reforç* s'intenta emular el comportament dels animals davant del mètode d'entrenament per càstig i recompensa, on l'algoritme ajusta el seu comportament basant-se en l'aprovació o penalització de conductes passades a fi de maximitzar les recompenses.

Un clar exemple de les aplicacions de l'aprenentatge automàtic és la classificació entre correu brossa o no desitjat i correu important. L'algoritme s'entrena amb grans quantitats de correus,

on se li específica quins són brossa. A continuació, troba similituds (patrons) entre els que han rebut l'etiqueta de correu brossa, que poden contenir paraules com «oferta», «gratuït» i «descompte», i depenent de la presència d'aquests ítems i com estiguin relacionats en correus futurs pot determinar si es tracta d'un correu brossa o no. Aquests algoritmes no necessiten regles directes donades pel programador, només un objectiu, i el programa s'encarregarà de crear el seu propi model de què és un correu brossa. Notem que el sistema no ha de ser precís en un cent per cent dels casos, de forma que podem trobar un correu que considerem brossa entre el correu important o al revés. Quan això passa, el localitzem i indiquem que no està ben classificat, el sistema *aprèn* que els correus que continguin informació similar a aquella han d'anar a una altra safata, i modifica el seu comportament per a que, amb correus futurs es tingui en compte aquesta nova excepció. El sistema podria agafar com a indicador de correu brossa informació que nosaltres no hem ni tan sols considerat, com ara que els correus que tinguin entre 17 i 32 comes no són del nostre interès, o que els que no tenen assumpte al missatge sí que ho són. Quan no sabem de quina forma l'algorisme està modelitzant l'ordre que li hem donat, és quan parlem de les caixes negres que vèiem més amunt.

Si la ciència de dades s'encarrega d'extreure patrons útils i rellevants d'un conjunt de dades donat, l'aprenentatge automàtic s'encarrega d'analitzar-les i establir correlacions entre dades que consideri similars. Tenir una gran quantitat de dades és indispensable per a poder fer aquesta anàlisis, però també és fonamental que aquestes dades siguin de qualitat. Un model que aspiri a generalitzar correctament, necessita que les dades amb què se l'entreni continguin suficient informació rellevant pel problema en qüestió que es vulgui tractar.

Hi ha una gran varietat de tècniques d'aprenentatge automàtic —regressió, classificació, agrupació, xarxes neuronals, etc.—, tanmateix, no existeix una tècnica millor o pitjor que una altra, cada una serà més o menys adequada d'acord amb la tasca que es vulgui fer; i alguns sistemes també poden ser híbrids i combinar diverses tècniques. Una bona tècnica per a la predicció de gustos musicals serà aquella que produeixi una millor resposta en la recomanació de noves cançons entre els usuaris. Es podria establir quin sistema recomana millor coneixent el nombre de casos en què dona una sortida o una recomanació que no agrada. A aquest valor se l'anomena *error* i és molt útil per a escollir adequadament el sistema que es voldrà fer servir. Un error del 0,5% a l'hora de predir els gustos musicals és excel·lent, però seria insuficient pel sistema que hagi de fer circular un cotxe autònom: ningú no estaria disposat a pujar a un vehicle autònom que cometés un error de conducció una de cada dues-centes vegades.

Dins l'aprenentatge automàtic, com acabem de veure, existeixen diverses tècniques que serveixen per a cobrir una gran varietat d'aplicacions, però la que s'ha popularitzat més és la que fa servir xarxes neuronals. Aquestes xarxes estan formades per unitats de processament simples que interactuen entre si, utilitzant una xarxa molt profunda de milers de capes que està inspirada en el cervell humà. De la mateixa manera que els àtoms es combinen per formar molècules i aquestes per formar cèl·lules, les xarxes neuronals consten de conjunts d'entitats que mitjançant la interacció es poden disposar per a formar estructures cada vegada més complexes, capaces de processar grans quantitats de dades. Així, les dades entren i les respostes o decisions s'executen en forma de sortida (*output*). Per si sola, una neurona està limitada a funcions insignificants que restringeixen els comportaments que pot adoptar per a processar les dades. Ara bé, el sistema resultant de la interacció de suficients neurones connectades entre si, pot ser extremadament sofisticat. Cada neurona pot reaccionar d'una manera específica als senyals d'entrada, per això el comportament d'un sistema quedarà determinat en funció de com estiguin connectades les seves neurones.[9]

Un dels objectius del disseny de xarxes neuronals parteix del principi naturalista que pren la biologia com a model i copia el seu funcionament per a fabricar les aplicacions tecnològiques pràctiques (Coeckelbergh, 2021, p. 67). Es tracta d'inspirar-se en els sistemes biològics per tal d'obtenir millors sistemes d'IA. L'argument principal a favor d'aquesta concepció és que, a través de la simulació de nivells baixos de processament (subsimbòlics) de dades en un marc de xarxes neuronals, emergirà la intel·ligència, la consciència o el pensament.[10]

A diferència dels ordinadors convencionals que funcionen amb *unitats centrals de processament* o *CPU* (per l'acrònim en anglès) i que processen cada conjunt de dades una rere l'altra, les xarxes neuronals compten amb nombroses neurones on cada una d'elles pot processar dades per si sola —encara que sigui a un nivell molt baix—, fent possible que una xarxa de neurones interconnectades pugui processar immenses quantitats de forma simultània. A diferència d'una CPU, les xarxes neuronals permeten que l'emmagatzematge i el processament de les dades no estigui separat, la qual cosa permet que la seva manera de gestionar les dades guardi una sorprenent similitud a la forma com ho fa el cervell humà. Això significa que, en les

[9] Això no significa que conèixer la disposició de les connexions ens permeti saber perquè es comporta d'una manera o una altra, com és el cas de les mencionades caixes negres.

[10] Aquest argument és el supòsit sobre el qual s'alça la concepció emergenista de la ment, que més endavant tractarem amb detall.

xarxes neuronals, que els processos de memòria i processament es realitzen paral·lelament, un mecanisme que els dota d'una gran eficiència i flexibilitat.

No obstant això, encara que molts professionals del sector es vegin enlluernats per formes tan antropogèniques de processar les dades, no es tracta en absolut d'una còpia computacional del cervell humà. El més rellevant d'aquestes xarxes és la seva capacitat per *aprendre* de manera jerarquitzada: les dades són assimilades i processades per nivells de capes. En un primer nivell, es poden reconèixer conceptes molt concrets i simples (per exemple, paraules com gat, taula o cotxe), i en els següents nivells, és fa servir la informació prèvia per a formar conceptes més abstractes i complexes (per exemple, frases com «el gat està sota la taula»). Quantes més capes tenen les xarxes neuronals, més complexos són els conceptes que pot assimilar, i d'aquí que aquests algoritmes també siguin coneguts amb el nom d'aprenentatge profund (que fa referència a la profunditat dels nivells de les capes).

La profunditat de les xarxes complexes permet al sistema fer abstraccions sense un entrenament previ, a diferència d'altres tècniques d'aprenentatge automàtic que sí que en requereixen. Són sistemes als quals se'ls fixen els objectius a complir i se'ls programa perquè els puguin assolir des de zero, sense *experiència* prèvia. Llavors, depenent de l'èxit de les respostes (sortides) que donen a les diverses entrades que reben, aniran modificant el seu comportament per acomplir amb l'objectiu assignat. A causa d'aquesta particular forma de processar la informació, l'aprenentatge profund fa que sigui molt complicat conèixer el procés a través del qual un programa que rep una entrada *A* executa una sortida *B*. És en aquest cas quan es considera que el sistema és una caixa negra; un sistema que sabem que funciona, però no sabem de quina manera ho fa. En efecte, encara que els programadors i desenvolupadors coneguin l'arquitectura del sistema neuronal, és problemàtic descobrir què passa a les capes intermèdies i de per quin motiu l'algoritme dona una resposta o pren una decisió. Fent servir un llenguatge antropocèntric es pot establir l'analogia de que el procés de reflexió *racional* de l'algoritme queda ocult.

L'expert en xarxes neuronals artificials, Dean A. Pomerleau, va ser un dels pioners en el desenvolupament de la conducció autònoma de vehicles amb ALVINN (per les seves sigles en anglès *Autonomous Land Vehicle In a Neural Network*) i també un dels primers en comprovar els perills dels sistemes que funcionen com a caixes negres (Pomerleau, 1991). L'any 1991, Pomerleau estava fent grans avanços ensenyant a una xarxa neuronal a conduir. S'asseia al volant d'un vehicle de l'exèrcit equipat amb càmeres i sensors mitjançant els quals un programa recollia tot el que passava a la carretera i les reaccions del conductor, a fi de modelitzar la forma

apropiada de conducció. Considerava que, amb un cert temps, el sistema hauria après les associacions necessàries per a conduir per si sol. Cada cop que es disposava a entrenar el programa, es posava al volant del cotxe i conduïa pels carrers de la base militar, deixant anar el volant de tant en tant per veure com reaccionava el sistema. Semblava que tot funcionava correctament fins que un dia, en una de les sessions d'entrenament, es va acostar a un pont i el cotxe va girar inesperadament cap a un dels marges. Afortunadament, va poder mantenir el control del vehicle i evitar l'accident. En tornar al seu laboratori, Pomerleau es va preguntar per la naturalesa de l'error que havia comès el seu sistema. Però, com saber-ho si havia programat la xarxa neuronal perquè aprengués a conduir creant el seu propi model de conducció? Era difícil descobrir en quin punt del procés d'aprenentatge el sistema havia pogut interpretar que un pont era una amenaça o un obstacle a evitar. No va ser fins que va analitzar minuciosament de quina manera responia l'algoritme a diversos estímuls visuals que va descobrir el que havia fallat. La xarxa neuronal havia utilitzat l'herba que creixia als vorals de la carretera com a indicador dels límits de la calçada i, per tant, com a referència de la direcció a seguir. D'aquesta forma, en anar a parar davant d'un pont, on no hi havia cap mena de referència en forma d'herba, el sistema va detectar que en aquella direcció s'acabava la carretera i que per això havia girar bruscament.

Per a desenvolupar un sistema basat en IA es poden fer servir o combinar diverses tècniques i mètodes segons la tasca específica que es vulgui que realitzi. La base fonamental d'aquests sistemes és el processament de dades en major i menor complexitat, permetent que els algoritmes d'IA puguin analitzar, aprendre i prendre decisions sobre aquestes dades. La capacitat per a operar i processar grans volums de dades és essencial pel funcionament efectiu dels sistemes basats en IA.

Podem categoritzar i avaluar les tecnologies basades en IA, segons el grau d'autonomia (entesa com a intervenció humana) que pot exhibir un sistema per a resoldre una tasca, en funció dels tres aspectes següents:

1) El grau d'orientació que necessita per a trobar patrons en les dades (supervisió, reforç o no supervisió): la forma en què els sistemes aprenen a identificar patrons pot variar segons la necessitat que les dades hagin de ser etiquetades.
2) La capacitat per a adaptar el seu comportament a partir de les dades recollides (retroalimentació): el grau en què un sistema pot aprendre i a ajustar-se a noves dades a mesura que són processades.
3) La capacitat per a operar amb dades dinàmiques i donar respostes adequades a diverses situacions (presa de decisió): l'autonomia que té un sistema per a prendre decisions

quan opera en entorns canviants per a respondre a circumstàncies noves de forma efectiva.

Com més complex sigui un sistema basat en IA, més complicat serà reconèixer per què està funcionant d'aquella manera. La complexitat d'un sistema es veurà incrementada en funció de l'autonomia que tingui i, per tant, com més autonomia tingui en els tres aspectes descrits en la categorització anterior, menys capacitat tindrem per saber per quin motiu dona una resposta.

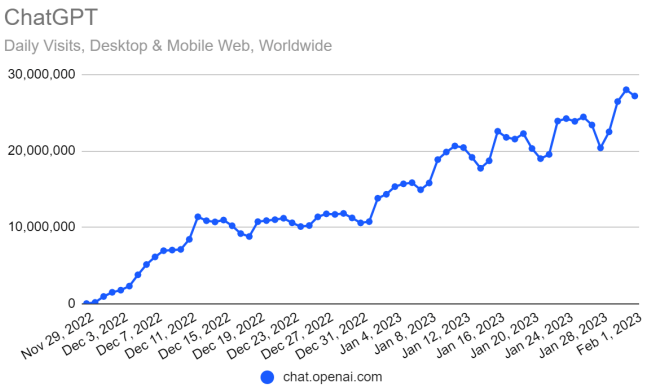
2.3 Algunes aplicacions i regulacions

Per fer una ontologia de la IA és important disposar d'una definició i una organització categorial, i també és central examinar les seves aplicacions i veure com s'està regulant. Ja hem dit que les aplicacions de la IA són molt diverses i depenent de cada aplicació pot requerir un nivell diferent d'autonomia. Una tecnologia es desenvolupa per tenir una aplicació en les nostres vides, tanmateix és llençada al món sense que sempre se'n pugui conèixer *a priori* quin ús se n'acabarà fent o en què derivarà. Encara que no sigui possible determinar els usos que tindrà la IA, és rellevant veure algunes de les seves aplicacions i de quina forma s'estan establint normatives per a regular-ne l'ús. És important entendre com s'està utilitzant avui per detectar l'impacte que aquesta tecnologia té en la vida —i el benestar— de les persones i corregir o limitar-ne els usos del demà. Com que la IA ha estat introduïda de forma salvatge i irreflexiva, és fonamental regular-la pels riscos que planteja en la invasió de la privacitat, les discriminacions per biaixos i la responsabilitat davant errors.

Les aplicacions dels sistemes basats en IA s'obren pas en camps com l'enginyeria, l'economia, el màrqueting, les relacions socials o, fins i tot, el sexe. Les recomanacions a través de les quals som guiats en les plataformes de compra en línia, funcionen amb tecnologia que utilitza IA i estan específicament dirigides als nostres interessos —modelitzats gràcies als historials de compra i cerca que deixem—, amb l'objectiu d'influir en les decisions de compra a través de la *publicitat segmentada*.[11] En l'àmbit de les xarxes socials, s'han creat programes informàtics que simulen ser persones, interactuant i relacionant-se amb la resta d'usuaris com ho faria un ésser humà, per què han après la manera en què ens comportem i relacionem. Són els *xatbots* o *bots conversacionals* i un dels quals ha captat més atenció és *ChatGPT*, que es pot fer servir lliurement des del 2022 quan l'empresa *OpenAI* el va llençar. Mostra de l'impacte que va tenir aquesta eina

[11] Consisteix en dividir al públic diana d'una campanya publicitària segons característiques preestablertes que responen als interessos del venedor —el que es coneix com a *target*—, a fi d'arribar als compradors potencials en el moment indicat.

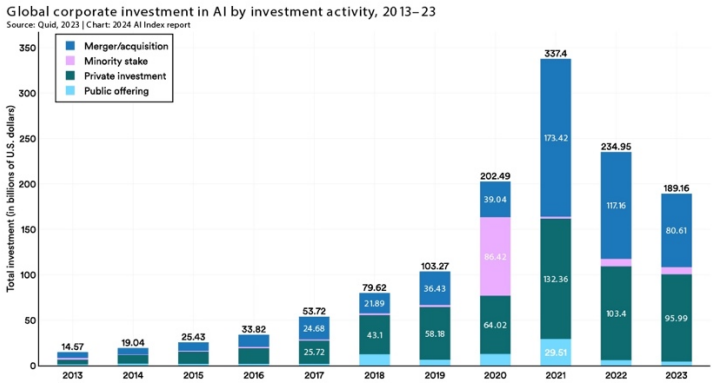
és que la primera prova gratuïta (del 30 de novembre de 2022) va aconseguir més d'un milió de registres en els cinc primers dies, una xifra que a principis de febrer del 2023 s'enfilava fins als gairebé trenta milions d'usuaris (veure Il·lustració 3: usuaris diaris de ChatGPT des del seu llançament).



Il·lustració 3: usuaris diaris de ChatGPT des del seu llançament

Font: (Barclays, 2023)

Les xifres confirmen que el model de llenguatge ChatGPT ha aconseguit desbordar l'interès dels consumidors des del seu llançament. De totes maneres, la inversió d'IA en àrees més àmplies s'ha anat generant durant els últims anys i no com a conseqüència dels últims desenvolupaments tecnològics, en tot cas la inversió corporativa ha estat la causa del sorgiment d'aquestes tecnologies innovadores (veure Il·lustració 4: inversió corporativa global en IA del 2013 al 2023).



Il·lustració 4: inversió corporativa global en IA del 2013 al 2023

Font: 2024 AI Index report

Si bé és cert que des de l'àmbit acadèmic les contribucions en el camp són notables, és en el sector industrial des d'on s'han destinat més esforços en la producció i millora d'eines basades

en IA. El motiu és, presumiblement, que les empreses poden veure reduïts els seus costos implementant tecnologies d'IA, un fet que suggereix que les diverses aplicacions estan ajudant a impulsar l'eficiència empresarial. La capacitat per a detectar patrons és molt beneficiosa en l'àmbit de la recerca científica, ja que pot ajudar a descobrir connexions que d'altra manera es tardaria molt més temps en detectar o es passarien per alt. A més a més, aquestes aplicacions també poden suposar un camí per proposar noves solucions a la crisi climàtica.[12] Gran part del mercat financer global està controlat per programes de *trading* automàtic, que analitzen constantment les dades per detectar petites variacions en l'economia, prendre decisions de compravenda i, així, maximitzar beneficis. Des de l'industria automobilística es fabriquen cotxes autònoms controlats per sistemes d'IA, i des del sector de transport públic —a poc a poc— també s'està incorporant aquesta tecnologia, que permet funcionar sense personal a bord.

En salut es fan servir els sistemes d'IA per poder analitzar les dades mèdiques dels pacients, per desenvolupar robots assistencials destinats a la cura de les persones i per fer diagnòstic, entre moltes altres aplicacions. Els sistemes d'IA mèdica estan demostrant un alt rendiment amb tecnologies innovadores com *EVEscape*, que vol ajudar a millor la predicció de pandèmies (Thadani et al., 2023), *SynthSR*, que és una eina que converteix les exploracions cerebrals clíniques en imatges ponderades d'alta resolució (Iglesias et al., 2023), o AlphaMissence, que està millorant la classificació de mutacions genètiques per detectar si són benignes o malignes (Cheng et al., 2023). En general, veiem que la IA s'està utilitzant cada cop més per a impulsar avenços mèdics, millorar l'atenció mèdica i desenvolupar dispositius que estan destinats a transformar el sistema de salut.

Tot i els cants de sirena que prediuen que la IA millorarà la productivitat de forma generalitzada, encara és aviat per preveure l'abast d'aquest impacte. Queden moltes preguntes per contestar i una preocupació important és què passarà amb els llocs de treball: en quina mesura s'automatitzaran llocs de treball en comparació amb els que —suposadament— es generaran amb la IA? No es tracta d'una qüestió menor, tenint en compte que les empreses estan integrant l'ús de sistemes d'IA de múltiples formes, però hi ha regions del món que estan experimentant més inversió que altres en aquestes tecnologies. Aquest fet planteja problemes

[12] Un grup de científics de la Universitat de Texas ha dissenyat un algoritme que, a través de xarxes neuronals, ha modificat un enzim per tal que sigui capaç de descompondre plàstics en dies, un procés que d'altra manera hauríem d'esperar segles perquè passés de forma natural (Lu et al., 2022).

de justícia donat que podria agreujar les diferències en innovació, productivitat i accés a les oportunitats derivades de la IA, a més pot dificultar la creació de marcs normatius.

La integració de noves tecnologies ha d'anar acompanyada del pertinent debat sobre l'abast del seu impacte. Malgrat això, molts cops, ens veiem obligats a acceptar que es desenvolupin i s'implementin sense poder formar part d'aquest debat, especialment perquè no es promouen la participació ciutadana en els afers tecnocientífics (Bastos et al., 2022). Encara que s'hagin fet nombrosos estudis per a conèixer millor l'opinió pública sobre les preocupacions i acceptabilitat que desperta la IA, són escasses les ocasions en què es convida a la ciutadania a formar part del debat de forma democràtica (de Fine Licht i de Fine Licht, 2020). Ha passat amb la introducció dels vehicles a motor, de l'enginyeria genètica, amb la producció d'energia nuclear i també està passant amb la IA. Ara bé, això no ens excusa a l'hora de plantejar-nos qüestions sobre com afectarà a les persones i a les diverses realitats socioculturals que conviuen amb aquestes tecnologies, i ho farà de forma diferent en funció del lloc que s'ocupi en la societat (per posar-ne algun exemple: estudiant, treballador, funcionari, empresari o polític). És important aclarir si polaritzarà encara més les desigualtats o afavorirà que s'anivellin; si podrà ajudar a les persones amb risc d'exclusió social a integrar-se o les vulnerabilitzarà encara més; si estarà al servei de tothom o només serà accessible per a uns quants privilegiats. En definitiva, s'ha d'estudiar a qui beneficiarà i a qui perjudicarà més, a fi de legislar de tal manera que esmorteeixi els danys i distribueixi equitativament riscos i beneficis. Si com a societat no som capaços d'entendre les amenaces i riscos que suposen determinats usos tecnològics davant els nostres drets, serà complicat reclamar-ne la regulació i, molt menys, esperar que els diferents òrgans legislatius se n'ocupin.

Per això cal que els òrgans reguladors i responsables polítics estiguin a l'alçada davant la necessitat d'establir regulacions per al desenvolupament i ús d'aquesta tecnologia. A dia d'avui està resultant complicat arribar a acords generalitzats sobre la legislació de la IA. Un document fet públic l'abril de 2021 per la Comissió Europea (European Commission, 2021)) es fa ressò d'aquestes preocupacions i explora diverses qüestions amb l'objectiu d'oferir un primer marc regulador de les tecnologies basades en IA. Un acord del desembre de 2023 entre el Parlament Europeu i el Consell de la Unió Europea, va establir la formulació de la Llei d'IA (European Parliament, 2024)—el primer marc jurídic a escala global sobre IA— que agafa com a eix vertebrador el document de la Comissió Europea. És una llei que va entrar en vigor l'agost de

2024 i que s'aplica a tots els sistemes d'IA que existeixin o estiguin en desenvolupament, però no serà d'obligatori compliment fins d'aquí dos anys, amb algunes excepcions.[13]

Paral·lelament, s'ha posat en marxa el Pacte d'IA (European Comission, 2024), que vol incentivar i ajudar a les organitzacions a aplicar de forma voluntària aquesta regulacions abans del termini legal. Amb la Llei d'IA es volen abordar els riscs associats als usos de les tecnologies basades en IA i posicionar a Europa per a tenir un paper de lideratge en iniciatives com aquestes a nivell mundial. La normativa té la finalitat d'impulsar la confiança en la IA, garantint la seguretat de les empreses i protegir els drets fonamentals de les persones. En la regulació s'estableix una classificació dels sistemes d'IA en quatre nivells segons el grau de risc que comporten: (1) *Risc inacceptable*, (2) *Alt risc*, (3) *Risc limitat* i (4) *Risc mínim*.

Al primer nivell (1), hi figuren els sistemes que suposen una *amenaça* per a la seguretat i que *vulneren* els drets fonamentals de les persones. Inclou sistemes que puguin fer servir per manipular el comportament humà, que incitin a la violència o fomentin conductes perilloses, i que puguin causar danys psicològics o físics. Tots els sistemes que comportin aquest nivell de risc hauran de ser prohibits. Serien il·legals sistemes de puntuació social com el que es descriu en el capítol *Caiguda en picat* (*Nosedive*) de la sèries distòpica *Black Mirror*, on l'estatus social de les persones queda determinat per l'habilitat dels individus per a obtenir bones puntuacions de la resta de persones en una aplicació.

Al segon nivell (2), s'hi inclouen els usos de sistemes que puguin suposar un alt risc per a la salut, la seguretat o els drets fonamentals de la ciutadania. Contempla tecnologies utilitzades en el transport (com la conducció autònoma), l'educació (com sistemes de qualificació automàtica d'exàmens) i l'àmbit de salut (com robots quirúrgics). En aquests sectors, els sistemes que s'utilitzin hauran de respondre a obligacions com l'anàlisi de riscs, la transparència i la traçabilitat de la cadena de presa de decisions, l'ús de dades de qualitat i la supervisió humana, entre altres.

[13] Per tal que les empreses es puguin familiaritzar amb la nova legislació i complir-la, les prohibicions es començaran a aplicar al cap de sis mesos, les normes de governança pels models d'IA d'ús general al cap de dotze mesos i les regulacions pels sistemes d'IA integrats en productes regulats al cap de trenta-sis mesos,.

El tercer nivell (3) considera sistemes com els xatbots, que només hauran de tenir un grau mínim de transparència i informar als usuaris que hi interactuïn que es tracta d'una bot conversacional i no d'una persona.

Per últim, el nivell més baix (4) suposa un risc mínim i, per tant, no s'especifica cap mesura obligatòria a complir pel seu ús. És el cas dels videojocs o filtres de correu. Tanmateix, la Comissió Europea anima a la creació de codis de conducta voluntaris per als sistemes d'IA de baix risc que queden recollits pel risc limitat i mínim (3,4).

La llei també especifica que els sistemes de reconeixement facial basats en IA (vigilància biomètrica massiva) tenen el seu propi nivell de risc entre la prohibició i l'alt risc. Per això, en concret, es prohibeix l'ús d'aquesta tecnologia en espais públics, alhora que s'està d'acord en consentir algunes excepcions. Això vol dir que les autoritats no poden fer servir els reconeixement facial per prevenir delictes o espiar possibles criminals, però se'n contempla l'ús quan sigui estrictament necessari: trobar víctimes potencials d'un crim, buscar infants desapareguts, prevenir atemptats terroristes, i la identificació localització, detecció o persecució d'un sospitós de delicte greu.

No deixa de sorprendre que, malgrat l'intent per regular els usos de la IA en sistemes de vigilància biomètrica, se n'accepta la utilització en situacions considerades excepcionals. El problema és que molts cops l'excepció no confirma la regla o, i igualment genera inquietuds com ara en cas d'error a l'hora d'identificar a un sospitós. Relacionat amb aquests aspectes tenim la utilització de sistemes d'IA amb finalitats militars, però en aquest cas directament no es fa menció a cap proposta de regulació. Ens ha d'alertar que no s'estigui estudiant posar línies vermelles als usos militars d'aquestes tecnologies, donat que són els que suposen una amenaça més gran per a les persones o, com a mínim, per a la població civil.

Les armes autònomes són equips que poden localitzar, seleccionar i atacar objectius humans sense que l'ésser humà participi directament en aquest procés, això no vol dir que no n'hagi de ser responsable. Sigui com sigui, dins el Programa Europeu de Desenvolupament Industrial en matèria de Defensa i l'Acció Preparatòria sobre Investigació en Defensa, la Comissió Europea està destinant 205 milions d'euros del seu pressupost en vehicles terrestres no tripulats, sistemes pe controlar eixams de drons i míssils d'alta precisió. Aplicacions tecnològiques com aquestes no haurien d'existir perquè degraden la vida humana d'acord a valors quantitatius. Però, es pot entendre —no justificar— que la seva existència està subjecta al fet que la vida humana resulta

cara a nivell militar i polític, mentre que és més desitjable tenir màquines que facin la guerra, estalviant baixes humanes en el propi bàndol.

En l'àmbit estatal, alguns països s'han ocupat d'aquesta qüestió fent polítiques sobre la regulació de la IA. La Xina, per exemple, ha introduït regulacions —normes que s'apliquen tant als proveïdors dels software com als usuaris— per abordar problemes relacionats amb les *deepfakes*[14], per prevenir que circuli contingut il·legal; per exigir el consentiment previ per a l'edició biomètrica. Els EUA han establert una legislació per adquirir eines de seguretat amb la finalitat de millorar la seva ciberseguretat, capacitant el Departament de Defensa estatunidenc per a adoptar tecnologies comercials innovadores amb aquesta finalitat. El Regne Unit ha anunciat la creació de l'Institut de Seguretat de la IA, la primera entitat amb recolzament governamental dedicada a promoure la seguretat d'aquesta tecnologia en benefici de l'interès públic i que té l'objectiu d'establir un marc sociotècnic per gestionar els riscs associats amb el desenvolupament i ús de la IA. En total, són 128 països que mencionen la IA en algun aspecte legislatiu, dels quals, 38 països han engegat algun projecte de llei realment relacionat amb la IA, una suma que representa 148 projectes de llei relacionats amb la IA (AI Index Steering Committee, 2024).

2.4 Conclusions sobre l'ontologia de la IA

Per a respondre a la pregunta sobre què és pròpiament la IA, hem tractat tres aspectes fonamentals: la definició, la categorització i les aplicacions i normes legislatives que la regulen. En síntesi, aquí podem concloure que la IA és una tecnologia que, basant-se en el processament de dades massives i la identificació de patrons, intenta realitzar tasques i resoldre problemes que normalment requeririen capacitats humanes.

El funcionament dels sistemes d'IA es basa principalment en seguir regles preestablertes per persones, per tant, entendrem que funciona algorítmicament i que poden *aprendre* de les dades que han analitzat i modificar el seu comportament. Per aquest motiu, molts cops són definits com a autònoms, malgrat que no poden fer res pel que no hagin estat programat. Depenent del grau d'autonomia (tècnica) que un algoritme tingui per a acomplir l'objectiu que se li hagi marcat, podrem identificar diversos sistemes d'IA, i classificar-los segons l'orientació humana que

[14] Tècnica basada en xarxes generatives antagòniques (*Generative Adversarial Networks*) que permet modificar imatges o àudios per a crear materials falsos on hi pugui aparèixer algú fent o dient coses que són mentida. Planteja seriosos riscos, ja que es pot utilitzar per difondre desinformació, manipular l'opinió pública o difamar persones amb continguts falsificats.

necessiten (entrenament); la capacitat per aprendre del seu comportament passat; i l'habilitat per a prendre decisions en entorns canviants. Quanta més autonomia presenti un sistema d'IA (respecte a orientació, adaptabilitat i presa de decisions), més complex és el seu funcionament, i més difícil és conèixer de quina manera està funcionant (com està modelitzant la realitat). És en aquesta manca de coneixement sobre el funcionament dels algoritmes d'IA que parlem de les caixes negres.

Les aplicacions dels sistemes d'IA poden integrar-se en múltiples àmbits, donades les seves diverses facetes millorant-ne la productivitat, però alhora plantegen preocupacions per qüestions econòmiques, polítiques i de justícia social. D'acord amb això, regular els usos de la IA és determinant per procurar garantir la seguretat, els drets fonamentals de les persones, la transparència en el desenvolupament i l'aplicació d'aquesta tecnologia en la societat.

CAPÍTOL 3. QUIN ÉS L'ESTATUS EPISTÈMIC DE LA IA?

Les múltiple aplicacions que es dona als sistemes d'IA s'insereixen per resoldre els problemes de les persones. Ens hem de plantejar, però, com aquestes tecnologies etnenen els nostres problemes i quines contribucions poden tenir fer al respecte. Un sistema d'IA pot tenir problemes? Realment pot resoldre els problemes que se li plantegen o només treu conclusions a través del processament de dades? Preguntes com aquestes ens orienten cap a la qüestió al voltant de l'estatus epistèmic de la IA, és a dir, com auquesta tecnologa capta la realitat i en quina mesura aquesta forma de captar-la pot tenir a veure amb la percepció i si podem considerar que constitueixi algun tipus de coneixement.

3.1 Les persones tenen problemes, la IA també?

Hi ha dos conceptes que ens ajuden a distingir els sistemes d'IA de la resta d'artefactes: l'autonomia i l'adaptabilitat. L'autonomia —fàctica o tècnica— els dota de la capacitat per a dur a terme tasques en entorns complexes i canviants sense orientació externa constant. Una capacitat que es contraposa a altres màquines que deuen la seva habilitat per a interactuar amb l'entorn a la de la persona que les condueix a través d'un control remot. L'adaptabilitat dels sistemes fa referència a la capacitat que tenen per a ajustar i millorar el seu comportament com a resposta a noves dades i entorns, similar al que les persones ens referim quan parlem d'experiència. Són dues competències —autonomia i adaptabilitat— que permeten als sistemes d'IA poder resoldre problemes sofisticats.

Entre els problemes que exigeixen una mecànica més complexa i avançada, tenim, per exemple, la tasca de la conducció autònoma. Els vehicles autònoms han de cercar i planificar la ruta més eficient per anar d'A a B (segons el criteri d'eficiència que se li hagi assignat), fer servir sensors que identifiquen possibles obstacles (programa de visió computacional) i ajustar la seva circulació davant la incertesa (prendre decisions). Quan diem que una programa de visió *entén* les imatges perquè és capaç de segmentar i classificar una representació dinàmica de la realitat segons els objectes que hi apareixen (vehicles, vianants, passos de zebra i semàfors), no ho fem referint-nos a que realment entengui les imatges: *entendre* no es pot reduir a un mer càlcul. De forma similar, no podem dir que un programa tingui problemes, sinó males modelitzacions, obstacles per superar i un patró de funcionament a seguir en diverses situacions. Seria imprudent intentar reduir el significat de problema a una explicació mecànica o funcionalista, la qüestió aquí és més aviat distingir què és un problema i què no ho és, a fi de veure en quina mesura els sistemes d'IA poden *comprendre* què és un problema.

Podem dir que trobar i formular problemes és un tret distintiu de l'ésser humà. De fet, una de les tasques de la filosofia, més que buscar respostes, és plantejar preguntes i descobrir nous problemes. Així, la filosofia es pot entendre com una disciplina encarregada de detectar problemes. Si el pensament filosòfic ha anat evolucionant —i, en principi, seguirà evolucionant mentre les societats existeixin— és perquè es detecten deficiències en els sistemes anteriors que convé superar i replantejar. Les noves preguntes obren les portes a nous problemes, els nous problemes exigeixen noves solucions i les noves solucions susciten noves preguntes. La reflexió sobre preguntar-se —fer-se preguntes— és essencial en l'actitud filosòfica i preguntes com *què és un problema?*, *per què ho és?*, *per a qui?*, etc., formen part d'una naturalesa humana que busca anar més enllà de la solució. Plantegem-nos quin podria ser el paper de la IA davant aquesta qüestió.

Una postura que s'ha adoptat des del sector tecnològic al voltant d'aquest tema, es rebel·la com un fonamentalisme tecnofílic que anuncia l'arribada d'una tecnologia que resoldrà tots els nostres problemes (representats per Elon Musk o Sam Altman, entre altres). Aquesta postura contrasta amb la seva contrapartida tecnofòbica, que profetitza el domini de la IA sobre els humans i l'origen de tots els problemes (com el *neoludisme*[15]). En un extrem hi trobem que el desenvolupament tecnològic ens alliberarà de tots els problemes —també dels problemes que generarà la mateixa tecnologia—, perquè la IA impulsarà l'economia fins al punt que les persones no hauran de treballar i perquè, aplicada a l'enginyeria genètica, trobarà solucions als problemes de la fràgil biologia humana. A l'altra extrem tenim els que defensen que el desenvolupament tecnològic desencadenarà en el tràgic escenari marcat pel domini de les màquines, on una tecnologia superintel·ligent acabarà sotmetent a l'espècie humana.[16] Si ens separem d'aquests punts de vista extrems, podem fer aportacions més valuoses per a l'anàlisi de l'estatus epistèmic dels sistemes d'IA que ens proposem.

Sabem que un dels trets característics de sistemes que utilitzen l'aprenentatge automàtic és revisar la idoneïtat de les respostes i aprendre dels mals resultats per corregir el seu comportament. Gràcies a aquesta funcionalitat de l'aprenentatge automàtic, comptem amb diverses computadores que, no només han estat capaces de derrotar als campions de jocs tan

[15] Moviment filosòfic que s'oposa al desenvolupament tecnològic i científic, que parteixen del llegat dels ludites britànics del segle XIX. Per més informació veure (Jones, 2006).

[16] Una enquesta que s'ha fet a més de quatre mil investigadors que es dedica a l'àmbit de la IA, ha revelat que la meitat d'aquests considera que existeix un 10% de probabilitat que els humans siguem destruïts per la IA (Stein-Perlman et al., 2022) i alguns experts opinen que aquesta és una situació que reclama més atenció que els efectes del canvi climàtic, donat que es podria donar abans (Tegmark, 2023).

complexes com els escacs o el Go, sinó que no trobaran mai un rival humà que les pugui superar. L'any 1997 el vigent campió mundial d'escacs, Garri Kaspàrov, va ser derrotat pel programa *Deep Blue* fabricat per la companyia IBM, en un matx a sis partides. Malgrat això, per ser justos, convindria destacar que no podem atribuir a la lleugera la victòria a Deep Blue, doncs Kaspàrov jugava amb desavantatge: vivía la fatiga física, mental i emocional que provoquen aquests enfrontaments en persones, al contrari que el seu rival digital. A diferència dels escacs, que a mesura que el joc avança hi ha menys peces al taulell i el ventall de moviments que un jugador pot fer se simplifica, en el Go la partida es va complicant més cada cop que s'afegeix una peça nova. Això és degut a que l'arbre de possibilitats que s'obre amb cada moviment, dificulta preveure les conseqüències de les jugades en el futur. *AplhaGo*, un software desenvolupat per Google, el 2016 va derrotar al Go a Lee Sedol (que aleshores tenia el segon lloc en títols internacionals) en una competició a cinc partides, deixant un balanç de quatre victòries de la màquina i només una per al seu adversari humà. El novembre del 2019, Sedol va fer públic que es retirava de la pràctica professional del Go al·legant que no tenia sentit seguir jugant perquè mai podria ser el millor del món, davant del que va catalogar com «una entitat que no pot ser derrotada» (Vincent, 2019).

Certament, aquest programa havia exhibit un domini del Go inassolible per a humans i actualment trobem algoritmes encara més sofisticats, com *AlphaZero*. Es tracta d'un programa que amb només tres dies d'entrenament va aconseguir derrotar per cent partides a zero a AlphaGo, el mateix que havia deixat sense opcions a Sedol al concedir-li una única victòria de cinc enfrontaments. AlphaZero no només és bo jugant al Go, també demostra una aptitud sorprenent per jugar a qualsevol joc, perquè no ha estat dissenyat per jugar específicament a cap joc en concret. Es tracta d'un sistema capaç d'assolir un alt grau de competència en qualsevol tasca perquè el seu comportament evoluciona en el temps, adaptant la seva resposta a les dades d'entrada.

Veient aquests exemples, ens podem fer a la idea de com els sistemes d'IA aborden la resolució de determinats problemes. També se'ns revela que la naturalesa dels problemes que poden ser resolts per la IA i els que afronten les persones són inherentment diferents. Mentre que un programa informàtic pot ser excel·lent per tractar problemes ben definits i estructurats, on les regles a seguir són precises i queden clarament delimitades, segurament no sigui l'idoni per resoldre problemes relacionats amb factors canviants i pluridimensionals o en contextos on les dades disponibles són escasses. Els algoritmes no operen bé sense dades i tampoc ho fan si no se'ls diu què és el que han de fer. En qualsevol cas, només són les persones les que tenen i

pateixen els problemes, perquè a banda de poder ser de càlcul o d'anàlisis (mecànics), aquests problemes també tenen a veure amb emocions, valors i prendre decisions (semàntics) que ens afectaran a nosaltres o a iguals. Per tant, un programa no té problemes, però pot ajudar a resoldre els problemes de les persones.

3.2 Què fa pròpiament la IA?

Si no podem ni tan sols igualar el nivell de joc d'una màquina computacional com Deep Blue, AlphaGo o AlphaZero, per què no fer com Sedol i deixar de jugar? La qüestió s'escapa del marc lúdic quan ens plantegem si té sentit continuar intentant resoldre problemes matemàtics si aviat la IA els podrà resoldre per nosaltres. Significa que aquests programes també són més intel·ligents que les persones? La gran competència que tenen els programes basats en IA per realitzar llargues cadenes de càlculs és incomparable a la humana. D'igual forma que totes aquelles tasques que exigeixen analitzar colossals quantitats de dades. Recordem que un dels propòsits de la IA és respondre als problemes que apareguin d'una forma similar a la humana, un intenció que destapa una presumpció que dona origen a aquesta disciplina: imitar el comportament humà és reproduir el pensament i la intel·ligència humans (McCarthy et al., 1955). D'aquí es deriva la suposició que exhibir una conducta intel·ligent és en si mateix *ser intel·ligent*, i una màquina suficientment complexa pot simular una conducta intel·ligent i, per tant, ser intel·ligent (Newell i Simon, 1976).[17] Alhora, la idea central que subjau és que la cognició humana no és altra cosa que un sistema de processament d'informació, de forma que un programa pot reproduir-la.[18] Es tracta d'una concepció algorítmica de la ment en la qual el seu funcionament es restringeix a seguir una cadena de regles i càlculs, sigui en algoritmes lògics, estadístics o en xarxes neuronals artificials complexes. Si respondre competentment (a la forma humana) a un problema donat és inequívocament signe de comprensió i, en conseqüència, de cognició o vida mental, haurem d'assumir que els sistemes d'IA que exhibeixin comportaments competents també són intel·ligents.

[17] La seva hipòtesi, que és a l'arrel del que s'anomenarà més tard *computacionalisme*, al·lega que la intel·ligència humana és un tipus d'operació simbòlica —donat que necessita d'un sistema de símbols—, per la qual cosa una màquina pot ser intel·ligent. Actualment, amb el desenvolupament de la IA, s'ha abandonat el supòsit que l'operació de sistemes simbòlics sigui responsable de la intel·ligència, per considerar que ho són les xarxes neuronals artificials que es fan servir en l'aprenentatge automàtic i l'aprenentatge profund.

[18] Un argument que ja era present en la filosofia de Leibniz amb el seu famós *Calculemus*, que redueix la deducció racional a un càlcul quan defensa que el raonament es podia fer tan tangible com una operació matemàtica, i davant d'una controvèrsia no faria falta discutir, sinó calcular (Leibniz, 1951, p. 51).

Es pot contestar a aquesta qüestió, com fa el filòsof de la ciència Daniel Dennett, apel·lant que hi ha mecanismes que escapen del nostre enteniment i que igualment cauen sota el nostre domini: no he d'entendre els processos cognitius que fan possible que puguem pensar per a poder pensar (Dennet, 2017). A aquest comportament Dennett el defineix com a *competència sense comprensió*, un enfocament que rebutja qualsevol classe de teleologia i defensa que tota competència (comportar-se amb destresa) és fruit de l'adaptació cega d'entitats a un entorn, sense cap mena de comprensió dels mecanismes que porten a terme.[19] Aquest argument permet explicar de quina manera els organismes s'obren pas en la vida mitjançant una voraç competència adaptativa que els permet sobreviure, sense la necessitat de comprendre de quina manera ho aconsegueixen. A través de l'evolució, les diverses espècies han estat dotades de recursos per a respondre adequadament al seu entorn: detectar el que és bo i distingir-lo del que és dolent, obtenir-ne un i evitar-ne l'altre i —tot això— ignorar el que no és útil per a la supervivència. Per Dennett, així és com funciona la selecció natural, sense més arquitecte ni finalitat que l'adaptació. Per un costat, del plantejament de Dennett se'n deriva que no estem justificats a atribuir intel·ligència, ni comprensió, a entitats que es comportin de forma competentment (ser no és semblar). Tanmateix, per l'altra costat, acceptar l'argument de la competència sense comprensió, més enllà d'una explicació a conductes adaptatives i extrapolar-lo les facultats de l'enteniment humà, també suposaria renunciar a la intencionalitat. En tal cas, hauríem de revisar si l'ètica i la moral no són més aviat mecanismes que ens ha conferit una evolució cega. De totes maneres, en el funcionalisme de Dennett —al qual tornarem més endavant—, podem trobar un element molt revelador per a tractar el tema de l'atribució d'estats mentals a sistemes en funció del seu comportament.

El pensament, la intel·ligència i la intencionalitat són nocions que defineixen l'ontologia humana i ens identifiquem amb elles sempre que ens (auto)definim. Degut a la necessitat de conceptualitzar la realitat en termes autoreferencials, també observem sota aquestes capacitats cognitives a la resta d'organismes i objectes del nostre voltant (plantes, animals i artefactes). Per aquesta raó, tendim a veure reflectides aquestes capacitats en entitats no humanes, interpretant els seus comportaments com si fossin la manifestació de les emocions, intencions o pensaments que tenen. Això és, antropomorfitzem el nostre entorn. Aquesta tendència es fa especialment evident quan ens fixem en el comportament dels animals, perquè els atribuïm motivacions

[19] En concret, Dennett es refereix per *competència* a la capacitat que té un organisme per aconseguir un fi (alimentar-se reproduir-se, sobreviure, etc.); i per *comprensió*, a una forma d'enteniment o habilitat de valorar i connectar idees per a explicar com funcionen les coses.

anàlogues a les que podríem tenir nosaltres si ens comportéssim d'aquella forma, encara que la seva conducta es pugui explicar des d'un punt de vista biològic o instintiu.

A l'Àfrica hi ha un tipus d'au, el drongo cuaforcat (*dicrurus adsimilis*), una espècie d'ocell que compta amb una estratègia extraordinària per aconseguir aliments (Flower et al., 2014). El drongo té l'habilitat de poder imitar tota mena de sons del seu entorn i, en la seva convivència amb altres espècies, ha après la forma que tenen els altres animals d'avisar-se davant amenaces potencials com ara depredadors o intrusos. Quan detecta algun perill, emet el so d'alarma perquè la resta d'animals amb que conviu puguin refugiar-se. D'aquesta manera, reforça la confiança del grup, que en sentir el so que informa d'un perill imminent saben que s'han d'amagar encara que no estiguin percebent res sospitós. El sorprenent és que utilitza aquesta informació en el seu favor. La seva tàctica consisteix en emetre el so de nou en constatar que algun membre de la seva comunitat de convivència ha aconseguit aliments, moment en el qual tots fugen atemorits i el drongo aprofita la confusió per apropiar-se del menjar oblidat. El comportament d'aquest ocell acostuma a qualificar-se d'enginyós, astut, intrèpid i un seguit d'adjectius que associem amb la intel·ligència humana.

Sota el focus de la consciència humana, aquestes aus estan posant la seva habilitat per a imitar sons al servei —si se'ns permet l'antropomorfisme— del seu *intel·lecte*. No obstant això, no estem justificats a defensar que aquesta competència que posseeix per a l'*engany* estigui relacionada amb al comprensió del seu comportament. Ben mirat, tampoc estem justificats a dir que es tracti d'un engany, ja que l'engany implica intencionalitat, i sense poder determinar si hi ha comprensió, es presenta igualment enigmàtica l'atribució d'intencionalitat. Una conducta adaptativa, per més sofisticada i eficaç que sigui, no ho és en virtut de la capacitat que tingui un organisme per a identificar-la, sinó en el fet que funcioni a l'hora de satisfer les seves necessitats bàsiques. Del contrari, seria com reconèixer astúcia en el color del pelatge d'un lleopard, que l'ajuda a camuflar-se i poder sorprendre a les seves preses.

Amb un propòsit similar, un estudi de l'equip liderat per Juliane Kaminski (2019), que investiga la cognició social comparada, sosté que l'expressió facial que posen els gossos quan volen menjar —i que assimilem fàcilment amb emocions humanes com la pena o la tristes—, no és altra cosa que un mecanisme evolutiu que els ha permès obtenir aliments dels humans. Els gossos domèstics han desenvolupat uns músculs facials que els permeten moure les celles, provocant una resposta afectiva i empàtica en les persones, que assumeixen que estan tristos perquè tenen gana i els genera l'impuls de voler alimentar a l'animal. Aquest comportament hauria donat un avantatge evolutiu als gossos que poden moure les celles respecte dels que no ho poden fer —

en forma de més aliments—, i hauria reforçat la presència d'aquesta capacitat en les generacions descendents. És cert que els gossos tenen gana. També és cert que mouen les celles per aconseguir menjar, un comportament que al llarg de les generacions ha resultat ser eficaç. Però d'aquí no podem derivar que el gos estigui trist o que sàpiga que ens farà pena i que, per tant, ens estigui manipulant en virtut de conèixer aquesta debilitat humana.

Gary Marcus, conegut per estudiar la intersecció entre la psicologia cognitiva, la neurociència i la IA, ha definit aquest fenomen amb la noció *escletxa de credulitat* (*gullibility gap*), que en contextos més generals es coneix com a pareidòlia[20]. És un fenomen que es dona quan els humans assumim com a intel·ligents comportaments aleatoris, només perquè encaixen amb preconcepcions que tenim integrades (Marchese, 2023). La convicció de que altres éssers es comporten de forma similar a la nostra, fa que projectem intel·ligència, comprensió i intencionalitat a les seves conductes. Marcus, exemplifica això a través dels models de llenguatge basats en IA generativa (com ChatGPT que hem vist anteriorment) (Marcus i Davis, 2020). Són sistemes que, encara que proporcionin respostes adequades (no vol dir que siguin verdaderes) a tot el que se'ls pregunti, no tenen comprensió de la realitat, perquè el seu funcionament consisteix a analitzar la forma en què les paraules estan relacionades (sintàcticament) i no el seu significat (semàntica).[21] Quan les persones utilitzem el llenguatge, vinculem les paraules a objectes. En canvi, els models de llenguatge actuals són estadístics, perquè de mitjana donen les paraules que estadísticament han anat seguides de les paraules que se'ls hi presenten. S'estableixen així relacions estadístiques entre les paraules (la freqüència en què apareixen juntes), sense que el sistema comprengui el significat que tenen. De la mateixa manera que parlar i escriure (utilitzar el llenguatge) no implica únicament saber establir correlacions estadístiques entre paraules, el mateix podem dir de ser intel·ligent, pensar o comprendre.

El fet que una entitat demostri un comportament que pot ser anàleg al que les persones tenen quan actuen intencionalment, no significa que es tracti d'una resposta també intencional. És la forma humana de conceptualitzar l'entorn —a la seva imatge i semblança— la responsable de precipitar la il·lusió de que la resta d'entitats comprenen per què es comporten com ho fan. El sistema que formula Dennett al voltant de la competència sense comprensió, ajuda a veure que és innecessari l'atribució de vida mental per a explicar els comportaments que observem. Que

[20] Percebre un estímul indefinit i aleatori com una forma reconeixible, per exemple, la figura d'una cara en un arbre, una roca o en qualsevol superfície.

[21] Tornarem sobre aquesta distinció en parlar de l'experiment mental de l'*habitació xinesa* que proposa John Searle per justificar que ser capaç de seguir rigorosament unes regles no implica entendre'n el significat.

el drongo ens sembli un astut lladre, els gossos hàbils xantatgistes emocionals o la naturalesa sàvia, obeeix al mateix mecanisme que ens fa pensar que un programa capaç de guanyar al campió de Go ha de ser també intel·ligent. Una actitud que ja queda delatada en el nostre llenguatge.

Quan diem que un algoritme ens recomana una cançó o que ens fa suggeriments de compres similars, ho diem de forma metafòrica, perquè realment no ens està fent cap recomanació ni suggeriment. Expressions com aquestes són les que fem servir en sentit figurat per a indicar la similitud entre les nostres accions i les funcions que acompleix l'algoritme, però en cap cas estem —o hauríem de creure que estem— davant de l'acció de recomanar o suggerir. La qüestió de fons és si, quan les persones fan recomanacions o suggeriments, estan recollint i comparant informació per a proporcionar la resposta —que serà algorítmica— que s'ajusti millor al model de preferències que han obtingut a partir de les eleccions passades d'un individu; o si estan fen alguna cosa més.

Defensar que realment recomanar no és només establir relacions estadístiques entre preferències passades i presents, també significa que no podem dir que el programa que fa funcionar la versió virtual del quatre en ratlla estigui pròpiament jugant. El programa no està jugant, tan sols nosaltres ho estem fent, doncs una de les característiques de jugar és gaudir, entretenir-se i aprendre. Per a què juguem i amb qui juguem, són les preguntes subjacents que ens porten a la vivència en un context relacional i amb certa igualtat de condicions. Si ens enfrontem a un sofisticat programa al qual no som capaços de derrotar, segurament deixarem de gaudir aquesta activitat i aviat no ens semblarà tan entretinguda ni divertida. El programa no estarà jugant, estarà anticipant millor que nosaltres el desenvolupament de la partida per assolir l'objectiu final (que en el cas del programa serà la victòria i en el nostre una vivència significativa, que mai podrà tenir la màquina). Això mateix és el que li va passar a Sedol. Si bé AlphaGo va guanyar el campionat, ho va fer sense jugar ni una sola partida. Per més incapaç que una persona sigui de derrotar a una IA en un joc, en realitat la IA no pot jugar.

No hem resolt l'assumpte al voltant de si els sistemes d'IA poden pensar o ser intel·ligents, sinó que hem aportat argument a favor de que determinar-ho no pot dependre únicament d'observar el seu comportament. El que sí que hem comprovat és que els programes basats en IA poden fer algunes tasques de forma més eficient i eficaç que els humans, gràcies a la capacitat que tenen per a processar grans quantitats de dades en un temps limitat i sense cansar-se. D'aquesta forma poden ajudar-nos a resoldre tota mena de problemes relacionats amb buscar informació determinada, la identificació de patrons i la predicció de tendències. Si tot el que són la

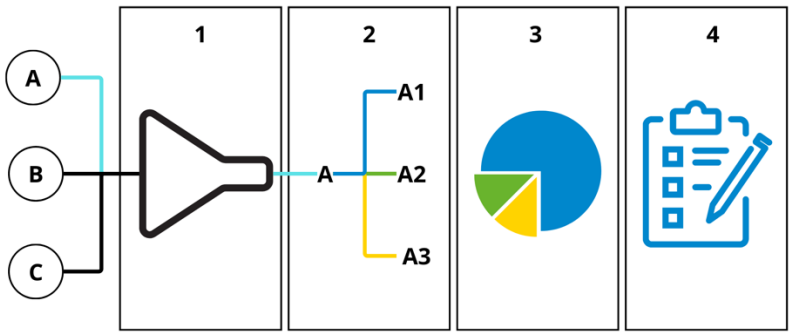
intel·ligència o el pensament és processament d'informació, com sosté l'analogia computacionalista del cervell-màquina, les persones no podríem actuar intel·ligentment o pensar davant la manca d'informació suficient.

3.3 Pot dubtar la IA? Una reflexió des de Descartes

Entre les activitats cognitives i definitòries del que és ser humà, una de les capacitats específiques és dubtar. Per Victòria Camps (2016, p. 25) «el dubte és el que ens constitueix, és el motor del canvi en tots els àmbits». Tant és així, que sense la capacitat de dubtar tampoc no podríem generar coneixement, ni tindríem aquests interès inesgotable per descobrir i entendre la realitat que ens envolta, no sabríem filosofar ni fer ciència. Sense dubte no ens qüestionaríem res, donant acríticament per vàlid tot el que se'ns presentés i sense fer possible que el coneixement evolucioni, la novetat o la creativitat. Ens fem preguntes perquè habitem el terreny de l'indeterminat: *què he de fer?, per què ho he de fer?, per què d'aquesta manera i no d'una altra?*

Fa prop de quatre segles que Descartes en el *Discurs del mètode* va establir les bases per a una anàlisi rigorosa i sistemàtica dels problemes, definint una sèrie de passos essencials per a avançar en el coneixement veraç (Descartes, 2010, p. 47):

1. No acceptar res com a cert fins haver reconegut clarament que ho és.
2. Descompondre el problema en principis bàsics més simples.
3. Elaborar noves veritats sobre aquests principis bàsics.
4. Reunir l'anàlisi de les parts i extreure un coneixement global sobre el problema inicial.



Il·lustració 5: diagrama mètode certesa Descartes

Font: elaboració pròpia

Descartes ens presenta un algoritme de resolució de problemes, concretament davant el problema del coneixement. En ell es proporcionen les regles descrites que, en seguir-les —en principi—, permeten la certesa suficient per aclarir tots els dubtes sobre el problema a tractar i generar coneixement (veure Il·lustració 5: diagrama mètode certesa Descartes). Així es planteja el dubte metòdic. El dubte és la peça central en el model cartesià i sobre aquest mètode s'articula la seva epistemologia, que culmina amb la famosa sentència «*Cogito, ergo sum*». Tan sols coneixem allò sobre el que no podem dubtar, això sí, abans hem de dubtar. El dubte ens ensenya a prendre distància d'allò que ens ve donat, a qüestionar-nos els propis prejudicis i a posar entre parèntesis el que se'ns presenta com a indubtable. Aquest mètode no suposa un dubte permanent i irresoluble —ja que llavors no quedaria espai pel coneixement—, sinó un dubte que ajuda a raonar millor. Seria possible programar un sistema d'IA que executés l'algoritme que descriu Descartes amb el seu mètode? És a dir, es pot programar quelcom semblant al dubte?

Inicialment, el mètode cartesià pot semblar desfasat[22] a l'hora d'abordar qüestions relacionades amb el desenvolupament de noves tecnologies intel·ligents —qüestions que reclamen certa urgència i pragmatisme en un món tecnològic que prioritza la rapidesa i l'eficàcia—. Tot i això, creiem que hi ha dos motius pels quals és una forma adient per a aproximar-nos als reptes epistemològics que es presenten en el context de la IA, uns reptes vinculats a què poden percebre de la realitat els sistemes d'IA i què poden conèixer (si és que poden conèixer). El primer motiu és que, segons el relat que defensa que els sistemes d'IA podrien tenir continguts mentals materialitzats per l'estructura lògica dels algoritmes, semblaria que mantenir un supòsit com aquest condueix a defensar una forma de dualisme extrem. Es tracta d'un dualisme que trasllada la dicotomia cos-ment al binomi hardware-software, limitant la ment a un seguit d'operacions mentals (logicoformals) que es poden reproduir en una base de silici. El segon motiu és que s'estan implementant els algoritmes basats en IA en molts contextos de presa de decisions, en entorns incerts i constantment canviants, on s'ha de treballar amb molt poques dades.

[22] En el sentit que hi ha un cert consens segons el qual el racionalisme ha estat superat per concepcions que, per exemple, es fixen en la importància de les emocions (Haidt, 2001), com també pel fet que històricament Descartes ha estat encasellat dins el dualisme metafísic (*res cogitans*/*res extensa*), un plantejament que es manifesta en el supòsit funcionalista de replicar la ment en una bases materials diverses. Malgrat això, Descartes afirma que, si volem ser rigorosos, només existeix una substància, que identifica amb Déu per evitar els problemes derivats del dualisme: «sols Déu és tal [substància] i no hi ha cap altra cosa creada que pugui existir un sol instant sense ser mantinguda i conservada pel seu poder» (Descartes, 1987, p. 52).

Les persones estan acostumades a treballar amb dades limitades i a prendre decisions en circumstàncies en les quals hi ha escassetat d'informació, però els algoritmes necessiten les dades per poder operar de forma eficaç. El mètode cartesià estableix que per arribar a la veritat és necessari revisar les preconcepcions a fi de despullar-les de prejudicis i obtenir així una base indubtable pel coneixement. En aquest sentit, podem trobar en aquest mètode un enfocament rigorós per avaluar les capacitats reals dels sistemes d'IA en contextos incerts. Quan s'aplica la tecnologia basada en IA entorns amb dades limitades, s'acostuma a confiar en models estadístics o de prediccions que poden estar molt allunyats de la realitat (un exemple l'hem vist amb ALVINN, el vehicle autònom de Pomerleau a l'apartat 2.2). El dubte cartesià és útil aquí per analitzar fins a quin punt els resultats, conclusions i decisions que generen els sistemes d'IA són fiables i en quina mesura estan condicionats per les limitacions inherents del sistema.

Si imaginem alguna cosa semblant a la *computadora cartesiana del dubte*[23], aquesta consistiria en un programa que, donada una entrada (*input*), sempre es qüestionaria la veracitat de les dades que rebés a través dels sensors que conformessin el seu *sistema sensitiu* i la forma com les processés el seu sistema operatiu. L'algoritme que fes funcionar aquesta computadora hauria de determinar què és veritat, atribuint aquesta etiqueta únicament a allò que no donés cabuda al dubte. Podria dubtar del correcte funcionament dels seus sensors quan detecten que és de nit, perquè les lents de les càmeres que fa servir estan brutes. Es podria plantejar si la seva capacitat per a modelitzar la realitat és fiable o si conté algun biaix quan, per exemple, infereix que en una imatge hi apareix un llop perquè està en un entorn nevat.[24]

I no només això, aquest programa no es limitaria a dubtar dels *inputs* o de la seva capacitat de processar-los, també hauria de dubtar sobre si ha estat programat de forma correcta i si tot el que capta i processa és un engany causat per una mala configuració. Encara que comptés amb demostracions lògiques i matemàtiques per a tots els problemes, que li vindrien donades amb el software i amb independència de les dades que pogués recollir amb els seus sensors, no hi podria confiar perquè podrien contenir errors. Així mateix, la computadora no es podria instal·lar permanentment en el dubte; hauria de desposar d'un algoritme capaç d'esgotar tota incertesa.

[23] Certament es té present que això suposa un exercici d'antropomorfització, però en aquest cas es pot justificar apel·lant que és un experiment mental per distingir el valor de dubtar per a les persones i la inoperabilitat que representaria el seu correlat computacional.

[24] Amb l'objectiu de veure els problemes de predicció que pot generar un caixa negre, un estudi (Tulio Ribeiro et al., 2016) analitza que el motiu pel qual un sistema de reconeixement d'imatge basat en aprenentatge automàtic confonia els llops amb els huskies era perquè havia après (amb les imatges d'entrenament) que els llops sempre estan acompanyats de neu.

De no ser així, suposaria que la computadora no podria fer res, quedant paralitzada com l'ase de Buridan, amb la diferència que en aquest cas seria davant la convicció que tot el que els seus sensors capten i el seu sistema processa és una il·lusió, caient en una espècie d'escepticisme radical. En acabar, què li quedaria per dubtar? Hi hauria alguna certesa a la que pogués arribar? Podria extreure algun coneixement de tot plegat? Anàlogament a la conclusió cartesiana «no puc dubtar que estic dubtant que dubto i que —en conseqüència— no puc dubtar que estic pensant»[25], la computadora podria adduir que l'única cosa que no pot dubtar és que està funcionant, tot i existiria la possibilitat que estigués funcionant malament.

L'experiment mental que acabem de veure serveix per il·lustrar la contradicció a què arribem quan s'intenta descriure el comportament d'entitats que no són humanes (antropomorfització) sota característiques humanes (el dubte). La forma d'operar dels sistemes d'IA és mou en el terreny de la lògica i l'estadística, on tot és expressable i calculable en termes de probabilitat. Un funcionament com aquest no pot garantir l'objectivitat (o neutralitat) ni l'exactitud de les seves respostes, donat que depèn sempre de la qualitat de les dades d'entrenament i de que l'algoritme no estigui esbiaixat; tot i que molts cops els sistemes d'IA es presenten com a eines que proporciona respostes objectivables i neutres. El raonament humà divergeix d'aquest funcionament, en la mesura en que integra experiències subjectives i intencions no sempre quantificables, i en especial pel fet que pot treballar amb l'ambigüitat. Mentre que els algoritmes tenen dificultats per gestionar l'ambivalència, ja que estan dissenyats per operar dins de models més o menys rígids i predefinits sobre els quals conceptualitzen la realitat de manera limitada; les persones ens sabem adaptar a la incertesa, en entorns permanentment canviants i sovint imprevisibles. Quan no troba més dades per a orientar el seu comportament, un algoritme respon amb inoperància, una persona respon prenen decisions.

Aquesta forma d'estar i entendre el món com un sistema dinàmic de relacions (amb un mateix, amb l'alteritat i amb el sentit que això comporta) juga un paper important, perquè permet que les persones puguin interpretar els diversos contextos —especialment els nous— molt millor que els sistemes d'IA. Entendre el context d'una situació donada és el que permet a les persones desenvolupar-se amb solvència, respondre a nous problemes amb certa rapidesa i prendre

[25] Respecte aquest argument, Bertrand Russell (2016, p. 219) considera que a través del cogito cartesià només podem conèixer l'existència indubtable dels pensaments i no la d'un jo pensant, de manera que dubtar que estic dubtant que dubto tan sols ens assabenta de l'existència del dubte. Tanmateix, encara que els pensaments no estigui produïts per nosaltres, algú n'ha de ser el destinatari —Frege considerava que pensar és la possessió de pensaments— i estem justificats a creure que el receptor dels pensaments pugui ser un subjecte pensant, que també és conscient d'aquest fet.

decisions amb escassa informació i sota ambigüitat. És en la pluralitat contextual on té sentit dubtar i sempre significa dubtar d'alguna cosa, per més que es pretengui dubtar completament de tot (Merleau-Ponty, 1993, p. 392). Descartes el que defensa és que aquest *dubtar* només s'esvaeix quan es té la certesa d'estar dubtant.

El punt crucial és aquí que la pluralitat contextual és inabastable pels algoritmes, perquè els seus models de realitat no donen espai a la complexitat i la variabilitat de situacions; situacions que no es poden compilar en un codi perquè un algoritme pugui anticipar com respondre-hi. El dubte existeix perquè interactuem amb contextos diversos davant els quals hem d'adaptar les nostres respostes per a actuar el més adequadament possible. Quan parlem de contextos, no només ens referim a un entorn, sinó també a la forma de ser d'un mateix on les alternatives són possibles i la vida és *indisponible* (Rosa, 2021).[26] Entrenar-nos per a prendre les millors decisions en cada situació demanar reflexió, i aquesta reflexió demana dubtar i arriscar-se (córrer riscos).

El dubte sorgeix perquè sempre que podem equivocar-nos respecte a alguna cosa, és en virtut del fet que també podem tenir raó (encertar) sobre aquesta.[27] Podem fer afirmacions errònies i estarem descrivint falsament la realitat, o podem fer afirmacions correctes i l'estarem descrivint de forma vertadera.[28] Encara que pretenguem dir la veritat, existeix sempre la possibilitat d'estar equivocats sobre què és veritat, podem no encertar en la forma com descrivim la realitat. Som éssers fal·libles, i ho som perquè les nostres pretensions de coneixement (què creiem que és veritat) estan subjectes a condicions d'èxit a les quals no sempre tenim accés (què és veritat). En vista d'això, les condicions d'èxit de les pretensions de coneixement poden ser desconegudes i aquest és un aspecte sobre el qual som fal·libles. Això no impedeix que el motiu pel qual fracassen algunes pretensions de coneixement puguin ser diversos. La nostra capacitat per a conèixer la realitat és essencialment limitada, revelant que no podem conèixer-la de forma absoluta, sinó de forma parcial perquè només tenim accés a una part d'aquesta. Per aquest motiu, no s'hauria de parlar pròpiament d'una veritat (absoluta) o d'un saber en si mateix. En comptes es pot parlar d'una pluralitat de formes de conèixer. El *pluralisme epistemològic* és un

[26] La IA es presenta com un càlcul absolut que domina totes les variables i Rosa ens dius que la vida és indisponible, fent referència a la idea que el món no pot ser completament controlat o dominat perquè sempre hi haurà una dimensió que escapa a la voluntat i manipulació.

[27] Tot i que aquest principi filosòfic universal es compleix generalment, existeixen circumstàncies en què, si més no, es podria posar en dubte, com ara quan provem d'analitzar si és possible equivocar-nos respecte a ser conscient.

[28] En una de les primeres definicions sobre la veracitat de les definicions, Aristòtil declara el següent: «és fals, en efecte, dir que el que és, no és, i que el que no és, és; verdader, que el que és, és, i que el que no és, no és» (Aristòtil, Metafísica IV. 7 1011b 26-28).

punt de vista que rebutja un únic sistema de creences i mètodes vàlids en el coneixement humà (Moulines, 1991, p. 30), donat que —tot i disposar de les mateixes capacitats cognitives— les persones poden construir multiplicitats de *representacions* coexistents de com són les coses.

Des de la tesi del pluralisme epistemològic podem entendre que trobem múltiples formes de tenir accés a la realitat, de tal manera que es pot compartir el mateix entorn i tenir experiències molt diverses de la mateixa realitat. D'acord amb això, no implica que estiguem accedint a realitats diferents, ja que la realitat sempre és la mateixa, el que varia és des d'on ens hi aproximem. Tampoc significa que quan percebem la realitat l'estiguem generant o construint i que no la puguem conèixer perquè els sentits a través dels quals tenim accés a ella la deformen, com es considera des del *constructivisme* (Hacking, 2000). Més aviat significa que quan percebem la realitat l'estem descobrint; no és que l'estiguem distorsionant a través dels nostres conceptes, sinó que la percebem de forma parcial. Que hi hagi una multiplicitat d'experiències de la realitat no s'ha de manifestar també forçosament en que hagin d'existir una multiplicitat de realitats.

En aquest punt és convenient tornar a parar atenció a l'exercici d'escepticisme que fa Descartes quan planteja que no existeix un dubte tal que pugui fer fracassar tota pretensió de coneixement. Si sempre ens equivoquéssim i mai poguéssim conèixer, també erraríem en afirmar que no podem conèixer i en que sempre ens equivoquem. Però, d'aquí també se segueix que no podem estar constantment qüestionant si sempre ens equivoquem, perquè aleshores tindríem raó respecte al fet de preguntar-nos si sempre ens equivoquem. Descartes conclou que sempre que ens preguntem sobre si podem estar equivocats —quan dubtem—, no podem errar sobre que ens estem plantejant si podríem estar equivocats —que estem dubtant—. Alguns autors (Gabriel, 2017; Kaufman, 2014; Schmaltz, 2019) han suggerit que l'argument cartesià, des d'aquest enfocament, convida a pensar en una forma plural de tenir accés a la coneixement, ja que proposa una epistemologia basada en regles contingents i més o menys provisionals (susceptibles de ser modificades). Això ens situa en un marc on no hi ha un ordre *a priori* de les coses[29], que queda de manifest amb la següent afirmació de Descartes: «quant a les opinions que jo havia acceptat com a vàlides, no podia fer altra cosa millor que procurar desfer-me d'elles, per a substituir-les llavors per altres de millors o per les mateixes un cop ajustades al nivell de la

[29] És l'expressió que fa servir Wittgenstein en la proposició 5.634 per a rebutjar l'essencialisme, entenent que les coses que observem en el món se'ns presenten de forma ordenada i coherent per poder ser reduïdes a un sistema de representacions, però que aquest sistema és contingent, donat que té alternatives possibles (Wittgenstein, 2012b, p. 165).

raó» (Descartes, 2010, p. 43). Amb tot això, la intenció de Descartes no és únicament posar-ho tot en dubte, perquè senzillament podríem estar equivocats al seu respecte; sinó perquè en dubtar s'evidencia l'experiència d'una pluralitat de formes de saber que no es poden fer desaparèixer, però tampoc unificar. No hi ha un coneixement absolut sobre la realitat que la representi de forma completa. Tanmateix, des del pluralisme epistemològic s'entén que sí que podem conèixer la realitat perquè es tenen diverses formes d'accedir-hi, encara que aquesta pluralitat de punts de vista (des d'on es mira el món) no pugui ser reduïda a una representació unificada.

Des d'aquest enfocament, el valor del dubte es troba en el fet que ens ajuda a distingir el que és vertader —i que suposadament esperaria ser descobert— del que és fals o il·lusori, obligant-nos a posar entre parèntesis les nostres creences, opinions i judicis. Dubtar serveix per avaluar les preconcepcions, arguments infundats i prejudicis a fi de destapar-los, eliminar-los i, en el millor dels casos, superar-los. Quan les persones dubten es posa de manifest la condició humana en tant que vulnerable, fal·lible i finita, permetent-nos pensar en els contextos i les conseqüències de les nostres accions. Està, per tant, al servei del procés de presa de decisions, per a adaptar la millor resposta a l'exigència de cada situació.

Situant-nos de nou en la qüestió sobre el dubte en el context de la IA, un algoritme està dissenyat per a treure conclusions davant les dades que se li presentin, i no té la capacitat d'interpretar aquestes dades, només de processar-les. Se li diu de quina manera ha de valorar les dades que rep perquè doni una resposta basada en patrons i dades prèvies. Per l'algoritme l'existència d'una veritat no té cap rellevància, busca proporcionar la resposta més adequada en funció de les dades disponibles i els condicionants per a acomplir una tasca. En general, un programa de computació només ha de poder fer aquelles tasques per a les quals ha estat programat. De manera que si fa alguna cosa que s'escapa de les seves funcionalitats és perquè està proporcionant mals resultats o funcionant incorrectament (Turing, 1947).

Posem per cas que li demanem a un informàtic que dissenyi un programa per comptabilitzar el catàleg de documents que hi ha a la Biblioteca de Catalunya per poder-ne consultar i organitzar l'estoc. Passats uns dies l'informàtic es presenta amb un programa que comptabilitza els llibres, manuscrits i diaris, però té moltes dificultats per classificar satisfactòriament altres documents com mapes, gravats, partitures, etc. Per més bé que el programa compti els llibres, manuscrits i diaris, el programa no estarà fent el que se li ha demanat perquè s'estarà oblidant de la resta de documents i, amb tota seguretat, comentarem a l'informàtic que aquest programa que ha fet no està funcionant correctament. Igual que el programa que no compleix correctament la seva

funció perquè ignora certs tipus de documents, un sistema d'IA que es basa en models estadístics pot fallar quan no representa correctament el segment de realitat que intenta modelitzar. Els models estadístics funcionen assignant probabilitats a diferents successos o resultats basant-se en dades anteriors, però aquests models només poden ser tan precisos com les dades i les regles amb les quals es basen.

La computadora cartesiana del dubte que ens hem imaginat tenia un objectiu de funcionament: dubtar de tot el que es pugui dubtar per a establir l'única cosa de la qual no es pot dubtar. En l'àmbit computacional, el dubte seria entès com una manera defectuosa de funcionar perquè, com hem vist, dubtar vol dir assumir que de la mateixa forma que podem encertar respecte alguna cosa, també podem estar equivocats sobre aquesta. Encara que es pot traduir en termes estadístics el grau d'èxit que tindrien un ventall de respostes (les possibilitats que aquestes siguin encertades), que una resposta sigui més o menys encertada depèn de factors indeterminables. Les persones dubten perquè han de prendre decisions sobre problemes que els afecten i que tenen conseqüències sobre les seves vides, i això ens revela dos aspectes que hem de tenir presents. Per un costat, que un algoritme no té problemes, com a molt, resol els problemes de —i per— les persones. I per l'altre, que decidir és una activitat lligada a la llibertat i, per tant, no és computable: una computadora no pren decisions, troba conclusions davant dades. Així, un sistema d'IA no pot estar basat en el coneixement, perquè només opera amb dades i la vida és més que dades.

La vida és recerca de sentit i de vida bona, que reclamen de judici i sensació. La vida transcorre en tres dimensions: la teleològica (direcció o objectius), la semàntica (significat) i la sensible (estatus ètic). Cap d'aquestes dimensiones es pot associar amb el que fan les màquines, robots, ordinadors o sistemes d'IA perquè els objectius venen determinats per la programació, operen amb símbols sense entendre'n el significat (com veurem amb l'habitació xines de Searle a l'apartat 4.1), i no tenen la capacitat de sentir. Un sistema d'IA només busca dades per a orientar els resultats que proporciona. Però, encara que pugui adaptar el seu comportament a les noves dades, és indiferent a la valoració dels resultats que dona.

El cas del dubte ens ha d'ajudar a veure que per poder establir alguna certesa, abans ens hem de qüestionar si allò que creiem saber és cert. Seguint-se d'això, podem dir que una computadora no pot construir cap certesa, perquè li manca la capacitat de dubtar (que no és reduïble a una probabilitat), i que, en conseqüència, el seu criteri per a la presa de decisions mai serà el coneixement o la veritat, sinó objectius donats. Si insistim a considerar que el funcionament dels algoritmes d'IA es basa en el coneixement i en una suposada capacitat per oferir les millors

respostes —l'autonomia els permet adaptar-se i l'adaptació els permet donar respostes que s'ajusten a cada circumstància—, estarem afavorint a fer que perpetuïn biaixos i errors sistemàtics. Un dels perills de considerar que als sistemes d'IA, com a les persones, els interessa el coneixement, és que no es contempli que les seves respostes puguin estar desviades dels nostres interessos per una mala modelització.

S'atribueix a Mark Twain una frase que resumeix el problema que estem comentant: «el que ens porta problemes no és el que sabem, sinó el que creiem amb certesa i resulta no ser així». Topem amb dificultats infranquejables quan donem per vàlid un argument infundat i ens neguem obstinadament a reconèixer la seva falta d'eficàcia. Però això no és degut únicament al fet que desconeixem la seva inoperància, sinó a què ni tan sols ens havíem plantejat (dubtat) que pogués ser erroni. El dubte ens ensenya a avaluar les nostres creences, a qüestionar el que creiem saber i a replantejar-nos la forma en què aquestes qüestions estan sent plantejades. Una altre fet que es vol demostrar amb l'experiment mental de la computadora cartesiana del dubte és que no tots els algoritmes són computables, mentre que qualsevol computació ha de ser necessàriament algorítmica.

Quant a l'apreciació que tota computació ha de ser necessàriament algorítmica, significa que que tots els processos computacionals han de seguir una sèrie de regles o instruccions. Respecte al fet que no tots els algoritmes poden ser computats per una màquina, no es pot dissenyar un programa que proporcioni una resposta per a tots els casos possibles perquè necessitarà temps il·limitats. Això ens porta a que, tot i que podem conceptualitzar un algoritme per abordar un problema, aquest no serà computable en tots els casos. Un exemple el trobem amb el *problema de la detenció* (*Halting Problem*), que explora la impossibilitat que, donat un programa i una entrada, es pugui determinar si el programa es pararà o continuarà funcionant per sempre (Strachey, 1965). Amb la computadora que hauria d'executar el dubte podem assumir que passaria el mateix, perquè no seria capaç d'avaluar totes les possibles situacions o resultats de les quals hauria de *dubtar*, culminant en un procés infinit de verificació i revisió, perquè el dubte (algorítmic) implica la constant reconsideració de les dades i dels resultats, un procés que mai arribaria a una conclusió definitiva.

3.4 Pot percebre la IA?

Gràcies als processos evolutius, cada ésser viu disposa d'una forma particular de captar i representar el seu entorn, amb un aparell sensor adaptat a les necessitats. De forma general podem dir que els animals compten amb un sistema sensorial per a representar la realitat, un

sistema que ha donat més importància a captar una informació que una altra, adaptant-se i sofisticant-se segons l'entorn. Els taurons i altres peixos cartilaginosos compten amb uns òrgans sensors anomenats *ampul·les de Lorenzini* que són sensibles als camps electromagnètics i gradients de temperatura per fer una representació del que tenen al voltant, a fi de detectar a altres animals, per exemple. Per orientar-se, els ratpenats utilitzen l'ecolocalització que els permet inspeccionar l'espai a través de sons d'alta freqüència i elaborar un model en tres dimensions del territori.

Les persones també experimentem aquestes variacions sensitives segons les necessitats entre integrants de la comunitat d'éssers humans. Quan es tracta d'una qüestió de supervivència, som capaços de distingir entre tonalitats del color blanc, com es diu dels esquimals (diferenciar on es pot trepitjar d'on no); desenvolupar l'oïda quan la nostra professió és la música (reconèixer amb facilitat notes); i quan, per exemple, les nostres capacitats visuals es veuen afectades, intensificar la sensibilitat del tacte per poder llegir en l'alfabet Braille (identificar les disposicions de punts).

El cas de Neil Harbisson és destacable en aquest context, ja que es va convertir en la primera persona en ser oficialment reconeguda com a cíborg. Degut a una alteració genètica (daltonisme acromàtic) que li impedeix veure els colors, Harbisson només pot veure en escala de grisos, i va idear una antena que l'habilita per a captar en forma de vibració sonora les freqüències de llum.[30] Testimonis com aquest ens demostren que la tecnològica ha permès augmentar la capacitat per a captar i representar la realitat, deixant al descobert espectres d'aquesta realitat que, d'altra forma, haguessin seguit ocults a l'enteniment humà. Amb el microscopi es van poder observar per primer cop els bacteris, les neurones i altres microorganismes. Amb el telescopi, Galileu va poder arribat a la conclusió que el sistema solar es devia comportar de forma similar a com ho fan els satèl·lits de Júpiter. I gràcies a l'espectroscopi, sabem de què estan compostos els cossos celestes. Totes aquestes eines han estat dissenyades per aguditzar les nostres capacitats perceptives i han obert la disponibilitat a *noves* realitats. Amb el nou panorama tecnològic, el desenvolupament de sistemes d'IA eixampla més aquesta obertura, fent possible reconèixer matisos i patrons en enormes quantitats de dades (establint relacions), assistint-nos en la presa de decisions (traient conclusions), i fins a proporcionar un model del pensament humà. Es fa

[30] El dispositiu que porta implantat permanentment dins el crani des del 2004 fa possible que Harbisson pugui escoltar els colors gràcies a un software que transposa les freqüències del color a freqüència sonora, descodificant el so de cada color, i fins i tot de percebre colors que l'ull humà no pot veure (Ronchi, 2009, p. 319).

visible així que el projecte d'idear una forma general de solucionar problemes (*General Problem Solver*), similar a la manera com ho fem les persones, és el Sant Grial de la IA (M. Boden, 2018).

Resoldre un problema donat en una situació concreta i en un temps limitat requereix contemplar una àmplia varietat de factors. Acabem de veure que hi ha mecanismes, com el dubte, que a les persones ens ajuda a prendre decisions, i no és computable, perquè involucraria processos algorítmics inacabables. Aquest és un fet que ha limitat el funcionament de la IA per a resoldre problemes a la forma humana, perquè el pensament humà té la capacitat de fer abstraccions (Flores, 2011). Conscients d'aquestes dificultats, els professionals del sector tecnològic han considerat que descriure i classificar mètodes mecànics per a formar abstraccions a partir de dades sensorials i d'altre tipus seria una empresa justificada (McCarthy et al., 1955), sota la premissa que el nivell d'abstracció en el processament de la informació és el camí per a modelitzar el pensament humà. Per fer un programa informàtic que pugui simular o imitar el pensament humà (o comportament intel·ligent) i aplicar-lo a la realitat, seria necessari que aquest programa disposés de coneixements sobre la realitat o, si més no, unes regles predefinides.

Complir amb aquest objectiu també involucra reflexionar al voltant de qüestions filosòfiques sobre què és el coneixement o de quina forma es pot obtenir, entre moltes altres. Davant les dificultats per a aproximar-se a un concepte d'intel·ligència humana, que és fruit d'una evolució biològica cega, s'ha optat per arribar a una definició conductual d'intel·ligència, que poc pot tenir a veure amb la naturalesa ontològica de l'ésser intel·ligent. Per aquest motiu els investigadors del camp de la IA han cregut que la manipulació lògica de símbols és una forma vàlida —consideraven que l'única disponible— d'enforcar el problema de la intel·ligència i el pensament humà amb el seu correlat computacional. Donades aquestes circumstàncies, es proposa (McCorduck, 1979, p. 264) que una entitat intel·ligent ha de poder resoldre quatre tipus de problemes:

1. Incorporar observacions específiques i generalitzacions a partir de les observacions.
2. Representar dades d'una altra realitat que la física (matemàtica, per exemple).
3. Adquirir informació sobre el món.
4. Assimilar i expressar internament aquests coneixements.

A partir d'aquí es pot extreure que serà intel·ligent tota entitat que pugui fer el seu model de realitat, respondre a qüestions sobre el seu model de realitat, ser capaç d'adquirir nova informació sobre la realitat i acomplir tasques d'acord amb els seus objectius aplicant aquests

coneixements. D'aquesta definició d'intel·ligència es deriva inevitablement una concepció representacional del coneixement (conèixer és fer representacions de la realitat), en la qual la informació sobre el món *exterior* —en aquest cas— és captada a partir de les entrades (dispositius sensors) i assimilada internament en forma de representació, sigui en una ment o en un ordinador.

A la llum de que les tasques que poden realitzar els sistemes d'IA tenen a veure amb el processament de les dades per a extreure, reconèixer i generar patrons, convindria plantejar-nos si el reconeixement de patrons constitueix una definició d'intel·ligència acceptable. Si realment es vol ser rigorós en aquest examen, també s'hauria de tenir en compte que els sistemes d'IA el que fan es calcular inferències estadístiques, de tal forma que no estableixen patrons exactes, sinó distribucions estadístiques de patrons. Per tant, hauríem de traslladar la pregunta al voltant de la intel·ligència als models estadístics i no a les màquines o ordinadors.

Recuperant la qüestió sobre la percepció, cal insistir en que va ser fonamental elaborar un model de la percepció va ser fonamental per a concebre la idea de xarxes neuronals artificials (McCulloch & Pitts, 1947). Originalment es van centrar en la percepció visual i auditiva, fragmentant les dades sensibles en ítems simples per processar-los com si es tractés d'una espècie de cadena de muntatge computacional. Amb l'objectiu d'automatitzar el reconeixement de patrons, Frank Rosenblatt (1957) va dissenyar el *perceptron* (abreviatura de l'anglès d'autòmat que percep i reconeix) que podia reconèixer lletres simples.[31] La percepció artificial és el propòsit que perseguien els primers professionals del camp de la IA, entenent-la com la capacitat que té un sistema no biològic (computadora, robot o altres eines tecnològiques) per a representar les dades que rep (reconeixement de patrons) d'una forma anàloga a la que s'assumeix que les persones utilitzen els seus sentits per a relacionar-se amb el món que els envolta (Turk, 2000). Mentre que la visió i l'oïda humana es poden cansar i estan limitades a un rang de freqüències, l'objectiu que persegueix el disseny de dispositius de percepció artificial és emular i superar les capacitats sensitives dels òrgans (sistemes) naturals, per a implementar-los en artefactes tecnològics o persones (com en el cas de Neil Harbisson) i millorar la seva eficiència en la realització de tasques.

[31] Un patró visual es registra com una impressió en una xarxa de neurones artificials que s'activen de forma coordinada amb la repetició d'imatges similars, activant una única neurona de sortida que respondrà amb un 1 si reconeix la imatge donada i amb un 0 si no ho fa.

Cal remarcar que per més que es millori la sensibilitat dels dispositius tecnològics per a captar qualsevol esdeveniment, això no significa que la màquina entengui el que està veient (píxels), escoltant (freqüències) o modelitzant (dades). La màquina no sap què és la realitat, de la mateixa manera que AlphaGo no sap jugar al Go —tot i ser imbatible— o ChatGPT no sap què és el llenguatge —tot i operar-hi amb absoluta normalitat—. La qüestió és si recollir i representar dades sobre el món pot ser considerat saber alguna cosa sobre la realitat. Al capdavall, no és el mateix ser sensible a detectar i representar estímuls, que el fet que la representació d'aquests estímuls constitueixi coneixement sobre alguna cosa.[32] La sensibilitat és la condició necessària de la percepció i no és una capacitat que estigui restringida a els éssers vius.

Entendrem que un objecte serà sensible a un estímul donat si, quan es presenta l'estímul, aquest objecte es comporta de forma diferent. La lent d'una càmera és sensible a la llum, un baròmetre és sensible a la pressió atmosfèrica, un termòmetre a la temperatura, un galvanòmetre al corrent elèctric i així una extensa varietat d'instruments científics. De forma separada, podem trobar dispositius artificials que superin la sensibilitat del seu correlatiu sistema biològic; ara bé, els éssers vius acostumen a ser sensibles a una major varietat d'estímuls, conjugant diferents sensibilitats, que qualsevol d'aquests dispositius artificials. Sigui per la major varietat d'estímuls que poden captar els éssers vius alhora, sigui per l'eina que fan servir per a processar-los (els cervells), hi ha una diferència essencialment qualitativa entre el que fa un sistema tecnològic en captar i representar un estímul i la percepció. Encara que l'assumpte al voltant de com es converteix un estímul captat en un símbol o representació (sigui en un cervell o en un software) és un problema que desafia el desenvolupament de la IA i l'estudi de la filosofia de la ment; creiem que a partir de l'anàlisi de la percepció, es poden fer importants aportacions en la discussió.

Tot i que no hi ha un consens generalitzable al voltant de la pregunta sobre la percepció, alguns autors (Goldstein, 2006; Merleau-Ponty, 1993) han suggerit que la percepció podria consistir en alguna cosa semblant a la reducció de la complexitat, ja que no es pot captar tota la realitat alhora. La percepció és conseqüència de la limitada naturalesa dels éssers, que impedeix poder percebre tots els estímuls disponibles de forma simultània i, en funció dels interessos, obliga a centrar l'atenció en un àmbit concret de la realitat. El que percep un organisme queda condicionat per les seves necessitats, interessos i motivacions —que són canviants—, i s'emmotllarà sota aquests criteris a l'hora de dirigir la seva atenció a un conjunt de dades i la

[32] Recordem la Il·lustració 2: diagrama relació dades, informació i coneixement.

forma d'interpretar-les. El coneixement previ (acumulació d'experiències) portarà a formar expectatives sobre el que pot passar en el futur i, al seu torn, el que s'espera que passi predisposa centrar l'atenció en una direcció o en una altra.

Hi ha diversos models del és la percepció, quins mecanismes intervenen en la configuració de les percepcions i de quina manera aquestes estan relacionades amb la realitat. Sumàriament podem repassar-ne els més destacats.

El model ecologista del psicòleg James Gibson (1972), que se centra en la percepció visual, concep la percepció com un procés simple on la informació es troba en l'estímul detectat, sense que sigui necessari un processament mental intern posterior. Gibson parteix del supòsit que en les lleis naturals de cada organisme descansen les claus *intel·lectuals* de la percepció com a mecanisme de supervivència, entenent que els organismes perceben únicament el que és valuós per acomplir amb les seves funcions vitals, o el que és el mateix: només és pot percebre allò que es pot aprendre i, alhora, és necessari per a sobreviure.

El model que proposa la psicologia cognitiva, molt influït per la teoria de la informació del matemàtic Claude E. Shannon i el biòleg Warren Weaver, reconeix que la interacció amb l'entorn no seria possible si no existís un flux constant d'informació (Hare, 1973). El flux constant d'informació és la percepció, que relaciona els estímuls que es reben, a través del sentits, amb l'entorn i amb els estats interns.

Des del conductisme, s'ha proposat que la percepció funcionaria més aviat com un procés que se situa en l'estímul i la resposta (conductual) que hi donem, contrastant amb el model de la psicologia cognitiva en entendre que els estats interns de la ment no tenen cap mena de rellevància, donat que tot l'anàlisi sobre la percepció es pot desenvolupar i resoldre des d'un punt de vista comportamental al nivell d'estímul-resposta (Skinner, 1985).

Aquests models de la percepció es fixen tots en que la percepció s'ha de dirigir cap a algun lloc, encara que discrepin respecte cap a quin lloc. Per bé que aquí no tenim l'objectiu de respondre a la qüestió referent a la direcció de la percepció de forma exhaustiva, a través de la distinció entre les estratègies de processament d'informació de baix a dalt (*Bottom-up*) i de dalt a baix (*Top-down*) podem explicar quins mecanismes estan involucrats en orientar aquesta direcció. En l'estratègia de baix a dalt es posa l'atenció en l'anàlisi de les dades d'entrada, sense que hi hagi regles *a priori*, de forma que és a partir de les noves dades que es representa (inductivament) el model de la realitat. El processament s'inicia en l'estímul detectat (a través del sistema sensorial)

i és a partir d'aquestes dades que es forma una percepció general, motiu pel qual aquesta estratègia també es coneix com a *processament guiat per les dades*. En el cas de l'estratègia de dalt a baix, es disposa de regles *a prior* per a orientar la modelització des de conceptes generals cap a aspectes simbòlics o concrets (deductivament). Els conceptes generals (estructures predefinides) són els que marquen la forma com s'han de processar les noves dades sensibles, per aquest motiu també es coneix com a *processament guiat per conceptes*. La principal diferència entre les dues estratègies de processament està en que la de baix a dalt es basa en les característiques sensibles de les dades, mentre que en la de dalt a baix és la rellevància de les dades el que guia el procés (Roediger & McDermott, 1993).

Generalment, els processos perceptius combinen les dues estratègies (de dalt a baix i de baix a dalt), atès que les dades que capta l'aparell sensorial són modulades per processos cognitius superiors. Que una de les dues adopti més o menys importància depèn de diversos factors (les dades disponibles, la qualitat de les dades, la durada de l'estímul i altres condicions ambientals): es desencadena el processament guiat per les dades quan les condicions ambientals són adequades i es poden recollir suficients dades, i s'afavoreix el processament guiat per conceptes quan les condicions ambientals no siguin idònies (Valdivieso & Macedi, 2018). Dit d'altra forma, en absència de dades òptimes, és processen els estímuls de dalt a baix, mecanisme que explica com es generen les expectatives i que afavoreix la predisposició a percebre un estímul concret (per exemple, quan escoltem el que esperem escoltar o veure el que esperem veure en funció dels nostres interessos, coneixements i experiències prèvies). Els prejudicis i les idees preconcebudes són la raó de la nostra forma de fer front a la incertesa, donant prioritat a percebre el que s'ajusti més al nostre model de la realitat.

Els sistemes d'IA simbòlica estan basats en l'estratègia de processament de dalt a baix i són els responsables dels *sistemes experts* —popularitzats durant els anys 70 i 80— que recullen regles pel reconeixement de patrons per resoldre problemes complexes i concrets (Leondes, 2001). Aquest enfocament és presentat com a IA *formalista* o GOFAI (per l'acrònim en anglès de *Good Old Fashioned Artificial Intelligence*) (Vellino, 1985)i és responsable de programes com *Mycin* (pel diagnòstic d'infeccions bacterianes) o *XCON* (per configurar sistemes informàtics basats en les necessitats dels clients). Els programes guiats per conceptes es basen en la idea que una representació ha de ser un símbol físic, localitzable en un cervell o en un xip. En canvi, l'enfocament *connexionista* de l'aprenentatge automàtic i les xarxes neuronals artificials es basa en l'estratègia de processament de baix a dalt, i entén les representacions en termes d'activació de conjunts de xarxes neuronals.

Si els sistemes experts són programes que han estat desenvolupats específicament per realitzar una tasca molt concreta, actualment amb les xarxes neuronals artificials es persegueix la mateixa finalitat sense que el sistema hagi d'estar programat específicament per fer una tasca en concret, sinó per dur a terme qualsevol tasca. En un procés que pretén imitar l'evolució biològica a través del mètode assaig i error, només es programa l'algoritme i se l'entrena amb grans volums de dades perquè modelitzi les seves pròpies regles per a realitzar la tasca en qüestió. Mentre que el funcionament dels sistemes basats en la IA formalista i simbòlica és transparent, perquè utilitza regles *a priori* i podem conèixer fàcilment els seus objectius, el funcionament dels sistemes d'IA basats en xarxes neuronals és opac perquè les regles —que ells mateixos construeixen— estan basades en induccions i generalitzacions de regularitats que no podem conèixer (Génova, 2023). No poder explicar les regles a través de les quals han modelitzat el seu comportament, fa que aquests sistemes funcionin com caixes negres, de les quals ja hem parlat prèviament i ens tornarem a ocupar per tractar les qüestions ètiques que involucren.

Per últim, —encara que no s'hi entrarà en detall— convé conèixer i destacar que hi ha dues conseqüències problemàtiques dels sistemes que utilitzen enfocaments de baix a dalt. Una és el sobreajustament, consistent en que, a mesura que el volum de dades creix en una tendència o patró de comportament, al sistema li costarà cada vegada més aprendre alguna cosa que no s'adapti a aquelles dades. L'altre és que les correlacions que els sistemes d'aprenentatge automàtic estableixen no impliquen causalitat; l'anàlisi i processament de dades no permet de forma autònoma (sense interpretació externa, per exemple) fer inferències causals, a fi de filtrar els patrons estadístics rellevants dels negligibles. Es traca de meres correlacions.

Per concloure, hem vist que la tecnologia ens permet percebre altres aspectes de la realitat, perquè hem estat capaços de fabricar artefactes sensibles (en un sentit mecànic) a estímuls que a nosaltres ens passen desapercebuts. Percebre no pot ser el que fan les diverses aplicacions de la IA, perquè té a veure amb la vivència i estar en el món. Els algoritmes no perceben absolutament res, sempre som les persones que percebem alguna cosa quan els utilitzem —integrant-los en sistemes— per a interpretar el processament de dades (sensibles) que fan, mostrant-nos altres complexitats de la realitat. Un sistemes d'IA, en tot cas, fa un model de percepció, però no percep res.

Els científics de la computació sostenen que la cognició humana es reflecteix en la capacitat per abstreure i aproximar patrons (Pasquinelli i Joler, 2021). L'abstracció és la capacitat d'extreure conceptualment una característica —essencial o accidental— del conjunt dels estímuls que captem a través dels sentits, i la IA és molt competent establint correlacions (reconeixement de

patrons). Tot i això, aquestes correlacions sempre són models estadístics i mecànics, i ens hem de plantejar fins a quin punt un model estadístic, si bé és capaç de separar les dades, representa el pensament o l'intel·lecte que intenta reproduir. Si per ser intel·ligent un sistema ha de complir els criteri de McCorduck (1979), trobarem que un termòstat serà intel·ligent perquè modela el seu entorn, representa la realitat en paràmetres de temperatura, rep constantment noves dades sobre el seu entorn i acompleix tasques d'acord amb aquesta informació (regular la temperatura). Aquest enfocament d'intel·ligència ha estat superat per un que l'encapsula en la capacitat d'extrapolar funcions més enllà de les dades conegudes.

Hem vist que hi ha dues formes d'aproximar-nos al problema de com es converteix un estímul donat en una representació, l'estratègia de dalt a baix i la de baix a dalt en el processament d'informació. Els sistemes d'IA poden fer representacions centrant-se en la deducció (representacions simbòliques) i en la inducció (representacions guiades per les dades). Mentre que la forma deductiva d'aproximar-se a la programació de les computadores requeria regles precises per a solucionar els problemes de forma aïllada —no per a molts dels problemes que es presenten en el la vida de les persones—, l'enfocament inductiu persegueix imitar un procés similar a l'evolució, descobrint la seva pròpia forma de captar la realitat i de representar-la.

És indiscutible que el pas d'un estratègia de processament a l'altra en el camp de la IA, ha permès el desenvolupament de programes tan sofisticats com ara ChatGPT, però no és en virtut de que el programa ara entengui millor el que escriu o se li diu. Aquest salt no deixa de ser quantitatiu, ja que ara els algoritmes disposen de colossals quantitats de dades (*Big data* o dades massives) a partir de les quals estableixen els seus patrons, prediccions i conclusions. No hem de deixar de veure el potencial que tenen les tecnologies basades en IA per a descobrir-nos nous coneixements sobre la realitat. Tot i això, no podem creure que les conclusions que proporciona estiguin basades en el coneixement, sinó en distribucions estadístiques de patrons mitjançant les quals s'estableix una funció matemàtica que es pugui assimilar a la comprensió humana. És en aquest sentit que la IA pot contribuir a les formes de comprendre humanes. Modelitzar la realitat no és ser intel·ligent.

3.5 El món de la vida, l'experiència i el cos

La percepció ens connecta amb el món i ens confronta amb un flux constant de dades (que se'ns presenta gairebé infinit) per poder dotar de sentit els conjunts de dades sensibles (objectes) en un entorn donat. A aquest entorn donat és al que Husserl es refereix quan parla de *camps de sentit*, la disposició en la qual els objectes estan relacionat. No es tracta de que amb la

percepció ens dirigim a objectes particulars aïllats que després situem en un entorn. Per Husserl és l'entorn o camp de sentit —i els objectes que conté— cap a on ens dirigim en la percepció (Husserl, 1980). En comptes d'un punt de vista privilegiat des d'on es pugui reconèixer i representar la realitat (en singular), existeixen una infinitat de camps de sentit —disposicions en les quals els objectes estan relacionat—, de manera que la realitat és una col·lecció d'esdeveniments irreductible a una única perspectiva.

Atenent a aquesta noció de percepció, un sistema d'aprenentatge automàtic pot captar (amb el reconeixement de patrons) propietats sensibles de la realitat i representar-les amb models estadístics, però sense arribar a percebre-la pròpiament, perquè és incapaç d'adonar-se de res fora —i a dins realment tampoc— dels seus límits lògics. La frontera epistemològica de la IA és l'esdeveniment anormal que escapa de la classificació i del patró reconegut (Goodfellow et al., 2014).

Considerem que és important evidenciar la diferència entre percebre —propi de les persones— , que té unes necessitats vinculades a la seva supervivència, del seu correlatiu model mecànic o estadístic, que busca satisfer unes ordres el compliment de les quals fixarà el seu —bon o mal— funcionament. Cal fixar-nos en la distinció entre el que percebem les persones i el que captens els sistemes d'IA, perquè el desenvolupament tecnològic persegueix el potencial dels sistemes d'IA per a prendre decisions que afecten a la vida humana. Es poden delegar a la lleugera aquest tipus de tasques a artefactes que en la seva modelització poden contenir fàcilment biaixos i representacions desalineades amb la vida i valors humans?

Amb aquest objectiu, aquí ens centrarem en els pensadors Aristòtil i Maurice Merleau-Ponty, ja que les seves filosofies, fins a cert punt podem dir que es complementen en l'abordatge de la percepció. Les seves perspectives filosòfiques situen el problema de la percepció al centre de les seves respectives reflexions i posen de rellevància que la percepció és un fenomen estrictament vivencial, irreductible a la suma de les dades recollides per l'aparell sensorial. El plantejament dels dos autors busca integrar les capacitats perceptives en un enfocament holístic que es preocupa per atendre degudament a una descripció del vivent, com un organisme que mai pot ser de reduït a un sistema de components *extra parts*.

El nexe d'unió entre Aristòtil i Merleau-Ponty es troba en la forma d'entendre la percepció, on hi ha un subjecte que es distingeix del món que el conté sense separar-se'n, en un intent per superar les descripcions —modernes— que confronten el subjecte amb un objecte. L'enfocament epistemològic dels dos autors coincideix alhora de perseguir una identificació

adequada dels actes perceptius, un punt de vista comú que segons el filòsof Renaud Barbaras[33] queda manifestat en què «la dificultat d'una filosofia de la sensació és arribar a pensar una unitat en què no es perdi la dualitat del que és subjectiu i el món» (Barbaras, 1998). Una identificació correcte de l'acte perceptiu ha de tenir en compte el compliment d'una doble condició on es dona una alteritat radical del subjecte respecte el món i, al mateix temps, un lligam inevitable entre subjecte i món. Tot acte perceptiu ha de complir la condició de *diferència* i d'*unitat*. La primera estableix la capacitat del subjecte perceptiu de distingir-se d'allò que està percebent. La segona diu que ha d'existir una comunicació originaria i indivisible entre el sentint (subjecte perceptiu) i allò sentit.

La concepció que Aristòtil té de la percepció no està exempta de controvèrsia i no permet fer lectures unívoques —o reduir-les a un sol enfocament—; com també passa amb altres aspectes del seu pensament, com l'hilemorfisme, la immortalitat de l'ànima o el primer motor. De forma resumida, podem destacar dues lectures sistemàticament confrontades sobre la figura de la percepció en el pensament aristotèlic, la dels historiadors de la filosofia occidental antiga Sorabji i Burnyeat (Caston, 2005).

Sorabji defensa que en Aristòtil la percepció ha de ser entesa com un procés fisiològic segons el qual l'òrgan sensible incorpora literalment la qualitat percebuda, una postura molt propera al fisicalisme, i segons la qual en veure objectes de color verd, els ulls d'alguna manera adquiririen el mateix color (Sorabji, 1992). Aquesta lectura manté que per Aristòtil la percepció és una forma d'assimilació on l'òrgan sensorial esdevé com el seu objecte.

Burnyeat no comparteix la idea que Aristòtil presenti la percepció com un procés fisiològic o un canvi material, sinó com un canvi o immutació de tipus intencional-espiritual —en termes tomistes—, de tal manera que en veure objectes de color verd, l'únic canvi que es produiria en el subjecte seria un *adonar-se* de que està veient quelcom de color verd (Burnyeat, 1995). Una interpretació de la percepció aristotèlica centrada en aquesta segona perspectiva ofereix més sortides cap a una comprensió vivencial de la percepció, que una centrada en l'essencialisme, i és en aquesta en la qual ens centrarem.

En virtut d'aquest plantejament, no és l'ànima la que percep, sinó la persona. Quan parla de percepció com d'un cert tipus de moviment, Aristòtil no s'està referint a cap tipus d'alteració,

[33] Considerat un dels fenomenòlegs més importants de l'actualitat i un dels estudiosos més autoritzats sobre l'obra de Merleau-Ponty, pensador a partir del qual generar la seva pròpia filosofia.

desplaçament o canvi material (moviments naturals), sinó a una particular forma d'activació d'una capacitat, poder o disposició que l'ànima posseeix (Sisko, 1996). Així, presenta l'argument segons el qual l'ànima no es pot moure per si mateixa, ho fa accidentalment quan el cos del qual és la forma es desplaça, encara que aquest moviment sigui motivat per ella. Com pot ser que l'ànima senti plaer, dolor, enuig o tristesa, afeccions que són moviment?[34] Si l'ànima no pot ser subjecte de cap moviment, què fa que pugui sentir? En realitat, per Aristòtil l'ànima no sent aquestes afeccions, més aviat és la persona que realitza tals accions «a través de» o «amb» la seva ànima que sent: «és millor no argumentar que l'ànima és compassiva, aprèn o pensa, sinó l'home en virtut de l'ànima» (Aristòtil, De anima, 408b 12-15). És més acurat dir que l'home —el compost de matèria i forma— és el subjecte que sent i pensa, i no una ànima de forma aïllada que, sota aquesta lectura, semblaria ser un atribut més del subjecte.

Quan ens endinsem en el pensament de Merleau-Ponty, també trobem una tensió en les interpretacions que se'n deriven, però en aquest cas és més pel seu propi filosofar de l'ambigüitat, en descriure la dualitat present en la unitat confosa de l'experiència. El seu punt de vista antropològic se centra en l'estudi del cos humà (estesiologia) com a animal de percepcions, entenent que és el cos —i no la consciència— qui percep la naturalesa que habita. No es pot iniciar aquesta anàlisi partint de la idea que un cos-instrument rep un pensament des d'un altre lloc. Tampoc es pot fer al revés, com si un objecte anomenat cos pogués produir misteriosament la consciència de si mateix. Segons Merleau-Ponty, no existeixen aquestes dues naturaleses —subordinades una a l'altra—, sinó un ésser doble o ésser de dues cares que sent i és sensible, veu i és vist, toca i és tocat (Merleau-Ponty, 2009, p. 205).

L'ésser doble és una idea neuràlgica en Merleau-Ponty, subjacent a tota la seva trajectòria filosòfica, que representa una concepció de l'ésser humà bidimensional i que es mou entre dos extrems, de forma anàloga al plantejament aristotèlic. En un dels extrems, trobem el positivisme que —resultat de l'herència del mecanicisme i l'empirisme— creu poder donar una explicació a tots els fenòmens naturals mitjançant relacions extrínseques entre els components físics o variables lineals tals com estímul-resposta. Davant aquesta tendència reduccionista fisicalista, Merleau-Ponty defensa que la vida (el centre d'accions vitals) i la consciència (el centre de significació) són irreductibles a una explicació mecanicista de l'ordre natural. A l'altre extrem, hi trobem l'idealisme, que a causa de les seves categoritzacions és excessivament rigorós a l'hora

[34] Convé especificar que per Aristòtil, per exemple, enfadar-se o espantar-se consisteixen en que el cor es mogui d'una o altra forma i que les afeccions esdevenen en virtut del desplaçament dels òrgans moguts o de l'alteració d'aquests.

de fixar els fenòmens i desatén la seva particular forma d'aparèixer. Crític amb l'empirisme, per captar de forma molt pobre els fenòmens, i amb l'idealisme, per sobrerepresentar-los, conjuga les dues cares d'una mateixa realitat (una associada amb l'ànima i l'altre amb el cos) en una unitat funcional i ontològica de la vivència humana: l'ésser doble.

No s'ha de confondre aquesta dualitat merleau-pontyniana de l'ésser amb una forma de dualisme (Ramírez, 2014, p. 23). El mateix Merleau-Ponty és conscient del paral·lelisme entre el seu argumentari i l'aristotèlic quan afirma que l'ànima pensa conforme el cos i no conforme a si mateixa, i conclou «s'ha de retornar a Aristòtil i l'Escolàstica, i concebre el pensament com a corporal, la qual cosa és inconcebible, però és l'única manera de formular davant l'enteniment la unió entre l'ànima i el cos» (Merleau-Ponty, 1986, p. 41). No s'està referint a que el pensament sigui corporal en un sentit estricte, sinó que si es vol ser fidel a l'hilemorfisme aristotèlic, aquesta és l'única forma de formular el problema.

La percepció és un aspecte central en Aristòtil i Merleau-Ponty perquè entenen que és la primera forma d'assimilació cognitiva, la forma primordial d'entendre la relació entre el cos i l'ànima, encetant el punt de contacte que forja la unitat entre la vida exterior i la vida interior. Hem vist que Aristòtil parla de la percepció com un tipus de moviment, en tant que és la particular forma d'activar les capacitats que l'ànima posseeix, si bé l'ànima és immòbil. També estableix que hi ha una correspondència entre la capacitat de moviment i la percepció (Aristòtil, De anima, II 1-2). No obstant això, no equipara la percepció amb el moviment. Més aviat, el moviment constitueix una dimensió física que acompanya a la percepció (un *amb*, *a través de* o *gràcies a*); una condició necessària perquè es produeixi la sensació, però no una condició suficient (Aristòtil, De anima, II 5). Per Aristòtil, el que manté en contacte el subjecte perceptiu i el món és una forma particular d'activitat senso-perceptiva i no un sentir o percebre pur, estàtic i aïllat.

En aquest punt, Merleau-Ponty també recorre i justifica aquest argument aristotèlic per mostrar que no existeix una forma de percebre les dades de forma aïllada, defensant que «el fenomen perceptiu està sempre en el context d'alguna cosa més; sempre forma part d'un camp [perceptiu]» (Merleau-Ponty, 1993, p. 26). L'idea de camp perceptiu (influïda pels camps de sentit de Husserl) condueix a Merleau-Ponty a parar atenció a les condicions d'indeterminació en les quals té lloc la nostra percepció. En l'acte perceptiu es deixen fora de la percepció formes abstractes i analítiques, no es perceben continguts purs i lliures d'ambigüitat. Per això, les nocions empiristes i idealistes de sensació fallen quan la conceptualitzen fora d'un context amb el seu significat vivencial. En la perspectiva merleau-pontyniana, la sensació és entesa en el

context d'un cos que viu enmig d'un món com a correlat primer de la seva activitat motora, una actitud que es pot veure representada quan escriu «ell [el cos] es veu veient, es toca tocant, és visible i sensible per a si mateix» (Merleau-Ponty, 1986, p. 16). Veure quelcom, és veure's a un mateix veient-ho en un mateix acte; percebre és percebre's a un mateix. Vist així, la percepció és entesa com un límit que, en la seva unicitat, obre dos plans diferents i alhora inseparables: el subjecte i el món. La percepció estableix simultàniament el contacte entre el meu ésser i el del món, ja que sempre que percebo, em percebo a mi mateix percebent alguna cosa. O en termes aristotèlics, la percepció implica una consciència perceptiva[35] que consisteix en *adonar-se de*, que acompanya cada alteració que es produeix en els sentits i que significa que aquesta alteració no passa inadvertida o desapercebuda (Aristòtil, Física, 244b15).

En definitiva, tant Aristòtil com Merleau-Ponty aborden la percepció com una activitat central per a la comprensió de la relació entre el cos i l'ànima o entre el subjecte i el món. La percepció no és la recepció passiva d'estímuls, ni es pot encapsular en una col·lecció de dades sensibles, sinó que combina una consciència perceptiva (percebre's a un mateix percebent la realitat) i una integració dinàmica amb l'entorn. Els processos perceptius es donen en un cos que es situa enmig de la complexitat del món.

En contrast, els sistemes d'IA operen principalment a través de càlculs estadístics i patrons —predefinits i autodefinits— als quals els manca la capacitat de percepció viva i contextualitzada. Mentre que la percepció humana és un procés integrador i bidimensional, la IA es limita a la manipulació de valors dins d'un marc purament matemàtic, sense accedir a la profunditat significativa i vivencial de l'experiència perceptiva. Així, les capacitats dels sistemes d'IA, per més sofisticades que se'ns puguin presentar, no poden captar ni substituir la complexa relació entre el subjecte i el món que es dona en la percepció humana. Això ens informa de que, tot i les habilitats que exhibeixin els artefactes tecnològics, hi ha una frontera epistemològica que no poden superar, una que la percepció humana travessa constantment.

A banda d'argumentar la incapacitat dels sistemes d'IA per a percebre, l'objectiu que ens proposàvem aquí també és evidenciar que ni les persones ni els nostres comportaments són un

[35] Una crítica que cal tenir en compte sobre aquesta consciència perceptiva de la qual ens parla Aristòtil és la de la regressió infinita, ja que si la percepció d'un mateix percebent que està percebent alguna cosa (consciència perceptiva) involucra un acte consecutiu, aleshores és necessària una nova percepció per a definir la presa de consciència de l'anterior percepció, i així successivament. Tanmateix, segons Caston (2002), aquesta característica de la percepció d'una percepció encaixa amb el sistema filosòfic aristotèlic presentat, que mostra el doble caràcter d'una percepció que unifica la dimensió que es dirigeix cap a allò sensible (món) i la que es dirigeix cap al sentint (subjecte).

patró algoritmitzable. Ha quedat patent que les tècniques de reconeixement facial classifiquen erròniament minories socials i que cometen discriminacions racials i de gènere de forma sistemàtica (Crawford i Paglen, 2021). La reflexió sobre com funciona la tecnologia, també ha d'estudiar com són els éssers humans (en plural i en singular), en especial, davant del fet que apunta Gabriel (Gabriel, 2019, p. 24) quan diu que «l'ésser humà és l'animal que rebutja la seva pròpia condició com a tal». L'ésser humà es rebel·la contra la seva naturalesa biològica, per intentar transcendir-la, com anuncien el posthumanisme i transhumanisme. Si es té en compte que la mateixa persona és nega a ser classificada: què pot significar una persona per a un sistema d'IA? Hem de tenir present que no es pot ensenyar a una computadora què és una persona sense reduir-la a una amalgama de correlacions estadístiques, i que això és un problema perquè no és el model estadístic el que constitueix al subjecte, més aviat és al revés. Els algoritmes processen les dades que generem per modelitzar i predir com ens comportarem, cosa que fan sense tenir en compte aspectes vivencials i intencionals que tenen a veure amb l'ésser humà.

La relació entre subjecte i realitat que ens proporcionen Aristòtil i Merleau-Ponty cobra sentit en tant que no ens trobem fora de la realitat que observem, tot al contrari, ens trobem immersos en ella sense poder-ne escapar. De la mateixa manera, encara que els models estadístics no siguin capaços de captar la realitat a la forma humana, no són artefactes passius que no hi intervinguin. Els dispositius basats en IA són llançats a la realitat, on intervenen en el *què* i el *com* fan les coses les persones i també en el que pensen.

3.6 Conclusions sobre l'estatus epistèmic de la IA

De la discussió al voltant de l'estatus epistemològic dels sistemes d'IA i la comparativa amb el de les persones, concloem reconeixentque l'adaptabilitat i l'autonomia dels sistemes d'IA els permet superar i resoldre problemes sofisticats. Tanmateix, donada la naturalesa dels problemes que poden resoldre (definits i mecànics), no funcionen bé quan hi ha falta de dades, quan aquestes són canviants, o quan no se'ls proporcionen regles (també els sistemes d'aprenentatge automàtic contenen regles en tant que han estat programats). En realitat, els artefactes tecnològics no tenen problemes (perquè no els afecten), malgrat que poden ajudar a resoldre els problemes de les persones. Que un programa informàtic guanyi a una persona als escacs o al Go, no vol dir que sigui més intel·ligent o que pensi; el programa no sap jugar perquè seguir un algoritme no és jugar (ni pensar ni ser intel·ligent). Juguem per a gaudir, entretenir-nos, aprendre: i és una forma de relació.

Quan interpretem que una màquina o un animal és intel·ligent perquè té un comportament amb el qual els humans demostrem intel·ligència, estem antropomorfitzant-lo. Aquesta és una actitud que associa injustificadament conductes similars amb intencionalitats similars. En aquest sentit, la competència sense comprensió ens revela que, o bé hem demostrat que la intel·ligència no requereix comprensió; o que encara no hem estat capaços d'identificar què és la comprensió en un sentit humà. El pas de sistemes d'IA simbòlics (de dalt a baix) a sistemes basats en la inferència (de baixa dalt) sofistica les tasques que poden realitzar i el seu grau d'autonomia. Però el seu funcionament se segueix basant en el processament de dades i el reconeixement de patrons estadístics (de grans quantitats de dades), sense cap mena de comprensió del que fan.

A més, decidir és una activitat lligada a la llibertat i, per tant, no computable. Un sistema d'aprenentatge automàtic no pren decisions, en tot cas troba conclusions davant dades i patrons estadístics. Dubtar ens ajuda a les persones a prendre millors decisions en circumstàncies incertes, sota ambigüitat, i arribar a algunes certeses; mentre que una computadora sota la manca de dades respon amb inoperància. Això significa que els sistemes d'IA no estan interessats en la veritat, sinó en proporcionar una resposta (encara que aquesta sigui inoperable).

Si entenem que la percepció consisteix en la reducció de la complexitat de la realitat, perquè no podem captar-la de forma simultània, els artefactes tecnològics ens ajuden en aquesta tasca ampliant la varietat d'estímuls que podem captar amb les correlacions que som capaços d'establir. Ara bé, som les persones les que percebem quan interpretem els models estadístics que ens proporcionen. La percepció vincula un subjecte amb el món (consciència perceptiva) i no es pot reduir a una recepció passiva d'estímuls sensibles. La tecnologia transforma la nostra forma de relacionar-nos amb la realitat, la forma en què la percebem i la forma en què ens autodefinim, si bé no sap què és la realitat ni què són les persones.

CAPÍTOL 4: QÜESTIONS SOBRE EL PROBLEMA MENT-COS

En aquest capítol volem contribuir en la discussió al voltant de la consciència, plantejant algunes de les principals concepcions que es presenten des de la neurociència i la filosofia, a fi de fonamentar si quelcom semblant a la consciència seria programable en un sistema d'IA. Si volem ser rigorosos en la identificació de què és la IA, hem de revisitar el problema ment-cos per tal d'advertir que les disposicions algorítmiques que processen les dades tenen poc a veure amb la subjectivitat que experimenta l'ésser conscient. Es tracta d'examinar, per una banda, l'estès supòsit segons el qual les persones són reduïbles al seu cervell, i, per l'altra, fins a quin punt les funcions que acompleix un cervell poden ser computables.

4.1 Sobre objectivitat i subjectivitat

Quan observem una obra d'art com el *Guernica* de Picasso, podem reconèixer els atributs objectius. Aquests atributs poden ser analitzats i processats per la IA. Es pot entrenar un sistema per a identificar i modelitzar els elements tècnics de la pintura, com l'ús dels colors, la composició, el traç del pinzell, etc. Un sistema de reconeixement d'imatges pot disposar d'una base de dades per a reconèixer patrons i establir-ne l'autoria. També pot examinar els elements que apareixen en la pintura per classificar-los: animals, persones i altres objectes, per exemple. Tanmateix, el conjunt d'emocions que es poden experimentar en contemplar l'obra d'art (malestar o el plaer, per exemple), i que respon als gustos propis de cada persona, són aspectes subjectius. Aquestes respostes emocionals i estètiques estan vinculades a la nostra experiència com a individus, les nostres preferències, el coneixement previ i la forma d'interpretar-les. A diferència del reconeixement d'imatges on es processen píxels (colors, formes, distribucions d'elements, etc.), les experiències subjectives no poden ser capturades ni replicades —més enllà del processament de dades i models estadístics— per la IA, perquè li manca la capacitat d'experimentar emocions i sensacions humanes de forma genuïna. Això destaca una diferència inherent entre el processament de dades que pot efectuar la IA i l'experiència humana. Tot i que la IA pot ajudar a identificar i analitzar característiques de les obres d'art, no pot forjar la riquesa de l'experiència personal ni la connexió emocional que desencadena en les persones.

Hi ha una creixent tendència *neurocentrista* que consisteix un intent per sintetitzar tot el que és l'ésser humà en un cervell i un sistema nerviós, en tant que el cervell és vist com el centre de processament dels éssers humans i el lloc on suposadament s'origina la consciència. Neurocentrisme és un terme encunyat per Markus Gabriel a l'hora de criticar la concepció que equipara el cervell amb la ment i l'autoconsciència i que assumeix que les persones som el nostre

cervell (Gabriel, 2016, p. 20).[36] Si volem entendre què és l'ésser humà, la ment, la consciència o el subjecte, no ho podrem fer a través de la filosofia, la sociologia o la reflexió introspectiva, sinó a través dels mètodes i eines que ens ofereix la neurociència per estudiar el cervell que podrem contestar a aquesta pregunta (Swaab, 2014). Des de la neurociència contemporània s'està alimentant aquesta ideologia neurocentrista, justificant que és el cervell el que ens fa funcionar i el lloc on podem trobar totes les respostes sobre què i qui som. Sense cervell aquestes funcions subjectives no es podrien dur a terme ni manifestar.

Ara bé, la qüestió formulada tal i com s'ha exposat sembla ignorar que a algú li agradi o li desagradi la pintura de Picasso, en darrera instància, depèn de les seves experiències, que són subjectives. D'acord amb això, podem entendre que s'argumenta fal·laçment dient que a qui li agrada el quadre és a un cervell, perquè aquest no té una existència aïllada fora del cos, i és a través del cos que podem experimentar. No és als cervells a qui els hi agrada el Guernico, sinó a les persones. No tindria sentit parlar d'una existència solipsista de l'ésser, protagonitzada per un cervell que aglutina totes les característiques amb què ens identifiquem. Menys encara, si tenim en compte que si som el que som és perquè vivim en societat i tenim la necessitat de comunicar-nos amb el nostre entorn.

És necessari tenir cervell per parlar de consciència, però no és una condició suficient per fer-ho. Comprendre el funcionament del cervell ens pot ajudar a conèixer què és la ment —de forma parcial—, però no a explicar-la de forma completa (si és que es pot explicar de forma completa o si pot tenir sentit voler-ho fer). Per un costat, el nostre cos està format per òrgans i teixits que estan fets d'altres tipus de cèl·lules que no són neurones, i de mecanismes que no s'expliquen a través de la neurofisiologia. I per l'altre, la interacció social és fonamental per a entendre el que som. Sense l'existència d'altres ments no tindríem la necessitat —no tampoc la possibilitat— de comunicar-nos i no hauríem inventat un llenguatge per cooperar i tenir més èxit en la supervivència. Tenir cervell no és suficient per explicar aquest comportament, és precís disposar d'una pluralitat de cervells i que aquests es combinin adequadament.

Per tant, la neurociència no pot contesta de forma absoluta preguntes sobre la conducta humana: per què som com som i què és la ment? Els mètodes i tècniques que recullen les neurociències són determinants per a explicar certs aspectes del funcionament del cervell, i la

[36]També són rellevants els conceptes neuromania i darwinitis als quals al·ludeix Raymond Tallis per a referir-se a aquest comportament que equipara tot el que són les persones al funcionament del cervell (Tallis, 2012).

seva contribució ajuda a entendre millor que és la vida mental, però no a explicar-la de forma completa. Es tracta d'un assumpte massa complex per explicar-lo exclusivament des del punt de vista de les neurociències i això es presenta com un argument en favor de què no som reduïbles al nostre cervell.

Entendre que existeixen entitats materials com els cervells i fenòmens subjectius com les ments és essencial si es vol explicar la singularitat de l'ésser humà. Es poden diferenciar objectes materials, tals com els cervells, i els processos que es duen a terme en ells, sense la necessitat de reduir-ne l'un a l'altre. Tots els cervells sans tenen un funcionament molt similar, però no hi ha ni una persona que sigui idèntica a una altra —o almenys tan similar com el funcionament dels seus cervells—. Ja alertava Hume sobre les dificultats per trobar el vincle causal entre la voluntat que mou el cos i el cos mogut per aquesta, preguntant-se: «hi ha en la naturalesa res més misteriós que la unió entre ànima i cos, en virtut de la qual [...] el pensament més refinat és capaç d'activar la matèria més grossera?» (Hume, 1980, p. 89).

Encara que els estudis neurocientífics proporcionin una comprensió cada vegada més precisa dels processos cerebrals, no capturen la totalitat de l'experiència humana (Rorty, 1983, p. 32). Les nostres ments van més enllà de la simple funció neuronal, són influïdes per factors com la cultura, l'educació, les experiències de vida i les pròpies interpretacions individuals (Popper, 1974). Hi ha una interacció complexa entre els objectes materials i mentals de la nostra existència, i és en aquesta relació material-mental on trobem la riquesa de l'ésser humà i la seva capacitat per experimentar el món de manera única i individual. És a dir, existeixen coses materials i una forma d'organitzar la matèria que possibilita les nostres ments. De la mateixa manera que certes configuracions materials permetrien que es desencadenin processos químics, la formació de galàxies i la vida, altres configuracions de la matèria possibiliten la ment.

Ens podem preguntar si la realitat és un fenomen independent de les observacions que se'n puguin fer. En aquest escenari, és convenient plantejar-se que la realitat existeix de forma objectiva per si sola (Schrödinger, 2016, p. 11). No obstant això, no es manifesta per la seva mera existència, ho fa condicionada pels processos que s'encarreguen de percebre-la. El pluralisme epistemològic explica que tenim diverses formes de descobrir la realitat —en cap cas de generar-la—, perquè hi ha multitud de formes de tenir accés a ella i, sobretot, perquè no la podem conèixer de forma absoluta. Aristòtil i Merleau-Ponty coincideixen, com altres autors, en que el vincle entre la subjectivitat i l'objectivitat s'explica com les dues cares de la mateixa moneda, perquè allò subjectiu i la realitat conformen una dualitat indivisible. El problema està en quina diferència hi ha entre els processos que descobreixen aquestes concepcions dels que no, quines

entitats materials es poden reconèixer a si mateixes reconeixent la realitat i quines no ho poden fer; en definitiva, es tracta d'entendre quins mecanismes cognitius i estructures internes permeten aquest reconeixement, i si hi ha una frontera clara entre éssers conscients i la resta.

En el marc de la IA, el problema ment-cos es (re)presenta de diverses maneres. En aquest cas, la relació queda corporitzada per un hardware, format pels components físics —electrònics, elèctrics i electromagnètics—, i un software, que inclou el conjunt de regles lògiques (algoritmes) necessaris pel funcionament del sistema. Quan considerem un sistema de reconeixement facial que utilitza el processament d'imatges, reconeixement de patrons, visió computacional i xarxes neuronals, podem distingir-ne els seus elements en aquesta divisió categorial. El hardware està compost per un suport físic que compta amb sensors d'imatge, processadors i memòria, que treballen conjuntament per captar, processar i analitzar les dades visuals per transformar-les en píxels i en informació significativa per a la identificació de rostres. El software és el conjunt de regles lògiques que permeten al sistema d'IA realitzar les tasques de reconeixement facial. Els sistemes algorítmics estan dissenyats per a extreure característiques facials, identificar patrons i treure conclusions en relació amb les coincidències —o dissemblances— referents a una base de dades de rostres. Podem afirmar que aquest sistema d'IA té un component físic (hardware) que realitza el processament de les dades visuals, i un component lògic (software) que utilitza algoritmes per a realitzar la tasca de reconeixement facial. De la mateixa manera que podem entendre que certes disposicions de la matèria produeixen la subjectivitat i aquest *adonar-se de* (unitat dual), s'ha d'examinar perquè això no és possible que es doni amb els components algorítmics i físics dels sistemes d'IA.

Per abordar aquest problema serà de rellevància explorar les perspectives filosòfiques tradicionals i també les consideracions científiques sobre el problema ment-cos. Per una banda, mitjançant l'ontologia que proposen el dualisme i el monisme es vol comprendre l'abast del problema sobre la ment, els diversos posicionaments filosòfics i les seves implicacions per a la comprensió de la ment humana. Per altra banda, un enfocament científic proporcionarà una comprensió detallada del funcionament del cervell. Obviar aquests estudis seria ignorar aspectes fonamental sobre el que sabem sobre el funcionament del cervell. El motiu és que sovint els resultats de la recerca científica, ajuda a corregir, matisar o repensar les aproximacions filosòfiques. Finalment, podrem argumentar fins a quin punt els sistemes basats en IA poden replicar o simular processos mentals i si aquest processar computacional i algorítmic pot ser equiparat a la forma com funciona un cervell o la ment.

El motiu pel qual hi ha una connexió entre allò mental i allò material és una qüestió que ha acompanyat de forma ininterrompuda a les tradicions filosòfiques, que han donat diferents respostes sobre aquest assumpte. Ment i matèria s'han desplegat sovint en l'imaginari humà com dues realitats contraposades que es defineixen per la mútua oposició. Propietats de la matèria tals com l'extensió, la posició i la densitat, són completament prescindibles per a explicar la ment. El problema ment-cos s'origina davant la qüestió sobre com es relacionen els successos corporals amb els mentals, el problema ontològic fonamental amb què lidia la filosofia de la ment. Les preguntes que enfronta el problema són principalment dues:

1) Com pot ser que els estímuls sensibles (físics) constitueixin l'experiència subjectiva?
2) Com és possible que el pensament (o voluntat) tingui influència en el cos?

Les respostes al voltant d'aquests interrogants demanen clarificar la necessitat del cos per a la ment i viceversa, una tasca gens simple de resoldre i que també suposa un repte per comprendre les idees de racionalitat i moralitat. Al cap i a la fi, subjau la pregunta: «l'ésser humà és racional o simplement una màquina biològica complexa?» (Metzinger, 2018, p. 262).

El problema ment-cos també evidencia la profunda separació que hi ha entre la filosofia que estudia la ment i les ciències que estudien la matèria. Mentre que la filosofia de la ment prové de la tradició analítica, que està continguda dins la mateixa branca que la filosofia del llenguatge i estudia els processos mentals i la consciència; les ciències físiques s'ocupen de conèixer i fonamentar la naturalesa de la matèria i la seva relació amb el món físic. Del problema de la relació entre matèria i ment se n'ocupa la recerca científica en branques com la neurociència, la psicologia cognitiva i, amb el desenvolupament tecnològic, les ciències de la computació (IA, aprenentatge automàtic i aprenentatge profund). Els enfocaments filosòfics i científics han desenvolupat marcs teòrics i mètodes diferents, dificultant la construcció d'una comprensió integrada de la relació entre la ment i el cos, així com de la naturalesa de la realitat.

Si el cervell és una màquina, i aquesta màquina té la capacitat de ser autoconscient en funció de com estiguin configurades les seves connexions neuronals, això significa que en imitar la seva configuració i estructures neuronals, també estarem reproduint la consciència? La mateixa pregunta és la que es fa Descartes. Les reflexions de Descartes sobre la dualitat entre cos i ànima no deixen de ser sorprenentment actuals quan anticipen el debat sobre si és reproduïble l'ésser humà, un debat que en aquests moments està centrant l'atenció en si la IA està en condicions de pensar de la manera com ho fan les persones o si podria arribar a ser conscient.

El filòsof francès considera que el cos dels animals és una màquina o autòmat creat per Déu i que es podria crear una còpia mecànica amb la mateixa disposició interna i externa que un mico (aspecte físic, arteries, músculs…) (Descartes, 2010, p. 80). Tal còpia seria indistingible del seu model de carn i ossos, perquè tindrien el mateix comportament i la mateixa aparença. Però, això no passaria si s'intentés fer el mateix amb un humà —copiant el seu aspecte intern, extern i comportament— perquè, afirma Descartes, sempre tindríem dos mitjans per a distingir el veritable humà de l'autòmat que el vol imitar. Primer tenim que l'autòmat amb forma humana mai podria fer servir les paraules ni altres signes equivalents per tal d'expressar el que pensa o respondre amb sentit al que se li hagi dit. I segon que per més que l'autòmat fes coses millor que les persones, inevitablement fallaria en fer-ne altres perquè no actuaria per coneixement a través de la raó —com fan els humans—, sinó per la disposició dels seus òrgans, que no podrien presentar disposicions tan diverses per enfrontar tota mena de situacions, com sí que permet fer-ho la raó.

Descartes s'està referint als mecanismes que permeten realitzar una acció, que en els animals venen condicionats per la disposició dels òrgans i les necessitats que tinguin (resposta instintiva), mentre que en els humans és la raó (resposta racional) què ho fa. S'assimila el funcionament del comportament dels animals amb el processament guiat per conceptes o de dalt a baix, on és necessita una programació específica per a realitzar comportaments diferents, i d'aquí l'impossibilitat de disposar tantes respostes com situacions possibles. Conclou la seva anàlisi declarant que l'ànima de les persones ha de ser de naturalesa completament diferent a la dels animals; i que allò reproduïble mecànicament són les funcions de l'ànima animal i el cos,[37] en cap cas l'ànima humana.

Descartes presenta així una de les primeres aproximacions al Test de Turing, en que el primer mitjà de verificació és superat actualment pels sistemes d'IA generativa (com ChatGPT), que poden fer textos i respondre pertinentment al que se'ls demana. Quedarà per discutir, tanmateix, si Descartes acceptaria això últim, donat que els models de llenguatge basats en IA modelitzen el pensament, però no són pensament.

[37] El dualisme cartesià estableix dues classes de substàncies, l'ànima i el cos, i alhora també distingeix dues classes d'ànima, l'animal i la humana, on l'ànima animal i la humana són naturaleses diferents de la mateixa substància. Tanmateix, s'entenen les incoherències i dificultats a què condueix considerar que l'ànima humana i l'animal són de naturalesa diferent, però també la mateixa substància, com sembla dir Descartes en alguns fragments.

La resposta del dualisme cartesià per explicar la tensió entre allò mental i allò corpori és que hi ha una distinció ontològica entre substàncies. Assumeix que l'ésser humà està compost per dues substàncies diferents, un cos i una ment. Postular la connexió entre dues substàncies de naturalesa diferent dona lloc al problema ment-cos, perquè és complicat justificar de quina forma aquestes dues substàncies s'afecten. És necessari explicar què és la ment, què és el cos i quin lligam s'estableix entre ment i cos.

Ara bé, és important clarificar que en Descartes el criteri que distingeix i caracteritza la ment és epistemològic i que no exigeix l'ontologia dualista. El concepte cartesià de ment —quelcom que pot existir independentment de la seva encarnació material—, agafa protagonisme a l'hora d'explicar els fenòmens mentals i, paradoxalment, establint una distinció ontològica innecessària per a defensar uns criteris epistemològics. En la Quarta Part del Discurs del mètode explica: «l'ànima per la qual jo soc el que soc, és enterament diferent del cos i fins i tot més fàcil de reconèixer que aquest, i, tot i que el cos no fos, l'ànima no deixaria de ser el que és» (Descartes, 2010, p. 60).

La característica primordial de la ment en Descartes és la immediatesa i infal·libilitat amb què és coneguda pel subjecte, és a dir, es tracta d'un criteri epistemològic. Enunciats psicològics tals com «tinc mal de cap» o «em pica l'esquena» són indubtablement vertaders per a un subjecte que descriu un contingut privat al qual té accés directe, sense la necessitat d'inferir que senti dolor o picor d'acord amb l'observació d'un comportament que així ho demostri. Però aquest punt de vista epistemològic pot admetre, sense entrar en contradicció, una ontologia que escapi del dualisme, perquè no obliga a admetre que el subjecte d'aquesta ment sigui una substància (Moya, 2006, p. 39). Es pot establir una diferència entre quan ens referim al dualisme de substàncies —existeixen dues substàncies—, de la concepció cartesiana de la ment, que és entesa com un àmbit de continguts i conceptes privats als quals tenim accés directe i infal·lible com a subjectes. Descartes, s'ha d'enfrontar al problema de com podem identificar les altres ments, ja que no tenim coneixement immediat i infal·lible de res més que dels estats mentals propis. Així doncs, si, tot i els inconvenients, estem disposat a assumir la concepció cartesiana de la ment, això no implica que haguem d'assumir també el dualisme ontològic.

Davant la desconnexió entre ciència i filosofia, Charles S. Peirce proposa elaborar una metafísica científica amb l'objectiu que les teories metafísiques utilitzin els coneixements de les matemàtiques i de la ciència per a construir els seus sistemes de pensament (Pierce, 1976), abandonant així la metafísica especulativa i les dificultats per a justificar-la. Per Peirce és un error considerar la metafísica com una ciència altament abstracta que està intrínsecament més enllà

de la cognició humana, un plantejament que és més propi de la teologia. En el seu rebuig d'aquesta idea, considera que la metafísica descansa en observacions i que és un saber que no s'ha de presentar com oposat al científic. Per això defensa una metafísica científica que es desenvolupi seguint el mètode científic i que es basi en l'observació. L'avantatge de la metafísica científica és que les seves hipòtesis es poden contrastar examinant si encaixen amb el coneixement científic actual. Malgrat això, podem apuntar que la proposta de Peirce sotmet la metafísica a un examen científic —per deixar de ser essencialment metafísica— i parteix de l'idea que la realitat només pot ser descoberta a través dels mètodes que disposi la ciència. En el seu abordatge, la metafísica tracta qüestions que van més enllà de l'observació empírica i que no es poden analitzar amb les mateixes regles que la ciència natural. Notem també que en aquest enfocament, la importància de la subjectivitat i la intencionalitat quedaria simplificada, perquè tan sols té importància la correspondència que les nostres pretensions de coneixement guardin amb les proves empíriques. Per tant, només es pot conèixer vertaderament tot allò que sigui accessible a ulls de la ciència, oblidant tot el que té a veure amb les relacions socials, culturals, històriques o altres sistemes conceptuals.

Una altra postura controvertida que intenta salvar l'abisme entre l'estudi de la ment i de la matèria és la que defensa la *metafísica digital*[38], que defensa que la realitat està composta per bits (1 i 0) —la unitat mínima d'informació— (Barrow et al., 2004; Chaitin, 2004). En primer lloc, manté que si les lleis de la física es poden expressar a través de programes informàtics, la matèria i també els processos que pot desencadenar, tals com la subjectivitat, són digitalment expressables amb bits. En segon lloc, la metafísica digital sosté que hi ha moltes raons per creure que el cervell humà és un ordinador universal. En conseqüència, simular el funcionament del cervell humà amb un ordinador ha de ser més una qüestió tècnica que una impossibilitat pràctica o funcional. Amb la capacitat suficient de processament no hi hauria d'haver cap impediment perquè un ordinador pogués experimentar la subjectivitat i l'autoconsciència, independentment de com funcionin els cervells. Aquesta doctrina veu l'univers fonamentalment com un sistema de processament d'informació, el coneixement com una col·lecció de dades, i les teories com a diferents disposicions d'aquestes dades.

És cert que les lleis de la física es puguin manifestar digitalment. Però d'aquí no es deriva que la matèria estigui composta per unitats de bits, ni que la realitat sigui absolutament computable. Respecte al fet que la realitat sigui computable amb bits, és una idea que xoca amb les

[38] També anomenada *filosofia digital* o *digitalisme*, no deixa ser com una nova versió del plantejament pitagòric que defensava que la realitat essencialment està composta per nombres.

conseqüències de la teoria de conjunts i nombres transfinits de Cantor, donat que, en principi, l'univers no conté suficient matèria o energia per poder fer una copia computacional de la realitat. Aquesta, és una qüestió que escapa de la nostra capacitat per a comprendre i de la de qualsevol ordinador (A. Moore, 2018). L'arrel del problema de la metafísica digital descansa en confondre les coses reals amb els models que tenim d'elles, perquè porta a concloure que la realitat està creada per la informació on «els bits són la informació que contenen els objectes físics» (Davies, 2004). Suggerir que la representació digital de la matèria és la matèria, és com afirmar que un mapa de Roma és Roma. Un mapa modelitza el territori, però no és el territori en tota la seva complexitat. En resum, la metafísica digital aconsegueix dissoldre el problema ment-cos, però ho fa a canvi de simplificar excessivament la realitat —fent-la indistingible d'un model sobre aquesta— i caient en un positivisme que no pot explicar res que no sigui definible a través de conjunts de bits.

4.2 Les respostes monistes

El dualisme substancial enceta una tradició de pensadors que situen l'existència de la ment separada de la cosa material, però no resol de quina manera estan connectades aquestes dues substàncies, ni tampoc la contribució d'una per a explicar l'altre. Considerar que existeixen entitats completament diferents com el cos i l'ànima, ha motivat la crítica de dues tradicions de pensament que, trobant-se ontològicament confrontades, han estat d'acord a assenyalar les deficiències de l'enfocament dualista.

El seu argument és que l'existència de dues entitats ontològicament diferents i separades no fa més que complicar la correcta identificació de la realitat: només hi pot haver una classe de substància fonamental. Les doctrines que assumeixen aquest punt de vista defensen el monisme, donat que entenen que tot el que composa la realitat es pot explicar a través d'una sola substància. Al respecte, el filòsof franciscà Guillem d'Occam considerava que les entitats no han de ser multiplicades sense necessitat, donant lloc al famós principi de *la navalla d'Occam* (Riesch, 2010). Aquest principi qüestiona pensar com a part del món entitats tals que el seu valor explicatiu sigui prescindible.

El monisme interpreta que la realitat està formada per un sol element que és la base de tot el que existeix, per tant, aquesta doctrina es caracteritza per atribuir a la realitat la condició de ser unitària. Arrenca amb la tradició presocràtica[39] i al llarg de la història s'ha presentat amb multitud

[39] Recordem que per Tales tot el que existeix són alteracions de l'aigua; que Anaximandre qüestiona a Tales dient que el substrat no pot ser res concret (llavors no es podria explicar la diversitat existent) i defineix el

de versions que podem agrupar en les següents formes de monisme: materialista, idealista i neutral. A la vegada, les tres formes de monisme també es divideixen en subclasses, en citem algunes: el materialista pot ser fisicalista (reduccionisme físic) o emergentista (el tot és més que la suma de les seves parts); l'idealista pot ser racionalista (només es pot conèixer a través de la ment-raó), empirista (l'experiència és un conjunt de representacions) o semiòtic (l'existència de les coses depèn d'un sistema de signes que es vol referir a elles); i el neutral pot ser general (matèria i ment són manifestacions d'una realitat subjacent) o energentista (tot el que hi ha és energia).

Els monistes idealistes situen el problema ment-cos en una mala interpretació de la realitat, entesa —erròniament— com una dualitat de substàncies i no com un tot. Consideren que tot l'univers és mental, immaterial i que les coses existeixen en tant que són observades o percebudes: *esse est percipi*.[40] El que anomenem *matèria* és simplement una col·lecció de percepcions o fenòmens mentals, i la seva existència es redueix a les experiències que en puguem tenir. D'aquesta manera, les lleis físiques i els objectes materials són vistos com construccions mentals, i no com entitats amb una realitat independent. Per tant, el seu supòsit parteix de la negació de la matèria, donant lloc al fenomenalisme segons el qual no podem conèixer la realitat en si mateixa, només els fenòmens que *apareixen* a la consciència, situant-se entre el racionalisme i l'empirisme. Aquests enfocaments condueix a una reconsideració del problema ment-cos, en tant que eliminen la distinció entre la ment com a subjecte perceptor i el cos com a objecte percebut, integrant tots dos en una única realitat (mental o fenomènica) que podem conèixer. Si l'idealisme transcendental kantià acusa al dualisme metafísic de ser responsable del problema ment-cos, de manera que superar aquesta concepció significarà superar el problema, el fenomenisme resolt declarant que tot el que no és mostra a la consciència és per principi incognoscible.

A causa de la seva posició radical i la seva distància respecte les concepcions majoritàries, l'idealisme ha estat criticat, en gran mesura, amb les mateixes objeccions que s'han presentat a la concepció cartesiana de la ment: la seva defensa facilita caure en un plantejament solipsista que no pot admetre l'existència de res que no sigui la pròpia ment, encarregada de generar una

principi com a *indeterminat* o *àpeiron* que és infinit, etern i muta constantment; i que Anaxímenes accepta les tres característiques de principi fonamental, rebutjant que s'hagi de tractar de res indeterminat i considera que ha ser alguna cosa que formi part de la naturalesa, al seu parer, l'aire.

[40] *Ésser és ser percebut* és la cèlebre frase del filòsof George Berkeley on es recolza la seva teoria immaterialista que va servir d'inspiració a Hume, Kant, Comte, Mach i als positivistes lògics, entre d'altres, per a rebutjar el concepte de matèria.

realitat aparent (o fenomènica). No obstant això, inspirant-se en una interpretació de la mecànica quàntica,[41] l'*idealisme metafísic* enuncia que la realitat no és independent de la consciència i que la ment és la que genera la il·lusió d'una realitat tangible i material (Kastrup, 2014).

Considerar que la condició de possibilitat d'ésser és ser percebut o observat i que, per tant, tot el que existeix és mental, és també incorrecte gramaticalment. Els verbs *observar* i *percebre* involucren una relació que s'estableix entre un ésser dotat d'un aparell sensorial i un objecte d'observació —incloent fenòmens i successos—. Si desapareix algun dels elements d'aquesta relació, la percepció o observació no tindrà lloc. De ser així, la consciència podria ser reduïda a la relació que hi ha entre un ésser —pensant— i les representacions que aquest fa de la realitat en aquets àmbit privat anomenat ment.

També seria una mala notícia per a les ciències naturals, que tan sols podrien explicar i descriure la realitat en termes autoreferencials de construccions mentals individuals i privades. La idea d'una realitat sense extensió sembla verdadera si argumentem que a) la ment no té extensió, i que b) hi ha una única substància existent, ja que aquest argument ens porta a concloure que c) la ment és la responsable de generar l'aparença d'un món material i tangible. Al cap i a la fi, semblaria un contrasentit preguntar-se sobre quant mesura o com de gran és la ment d'una persona. Ara bé, segons alguns autors el pressupòsit a) és fals —perquè no es proporciona cap demostració al seu favor— i acceptar-lo com a vertader, en realitat, és cometre una petició de principi, donat que s'assumeix com a vertader el que es vol demostrar (Double, 1999). Altres autors qüestionen si no s'estarà admetent massa ràpidament que la ment sigui una substància, i que fer-ho pugui ser un pressupòsit enganyós, perquè la ment veritablement és una propietat —entre moltes altres— del cos (Moya, 2006, p. 25).

La ment també podria ser una propietat del cos. Una propietat que en desaparèixer el cos, desapareixeria també amb ell. Per tant, equiparar la ment amb una substància seria un supòsit infundat perquè no pot existir independentment d'un cos. Mentre que considerar que és una propietat, vincula cos i ment alhora que els distingeix, i permet mantenir-se en el monisme, sense la necessitat de duplicar substàncies. A més a més, que la ment no tingui extensió tampoc seria una demostració en favor de l'idealisme, doncs la immaterialitat és un característica comuna de

[41] Una interpretació que parteix de la idea d'un univers no mecanicista, que planteja James Jeans arran dels nous descobriments en mecànica quàntica: «el curs del coneixement s'enfronta a una realitat no mecànica: l'Univers es comença a assemblar més a un gran pensament que a una màquina. La ment ha deixat de semblar un intrús accidental en el regne de la matèria» (Jeans, 1937, p. 188).

les propietats. La realitat sense extensió és equiparable a un software sense hardware: tenir hardware és condició de possibilitat de tenir un software, però no al revés, perquè el software necessita d'un hardware per poder operar de la forma amb què ha estat programat. I encara més si tenim en compte que qualsevol programari necessita estar materialitzat en algun suport físic per dissenyar-lo, fabricar-lo i executar-lo.

El monisme neutral explica la diferència entre matèria i ment com dues formes d'ordenació (Russell, 1982, p. 144). Defensa que tant els fenòmens materials com els mentals poden formular-se a partir d'un nivell subjacent que no és ni mental ni material, sinó *neutral*. Són les diverses formes d'ordenar o categoritzar la realitat les que donen lloc a conceptes que no tenen res a veure entre si (Stubenberg, 2008), com la matèria i la ment. De manera que matèria i ment són dues formes diferents de captar les organitzacions causals que configuren la realitat. Seria com si es volgués establir una classificació de les obres d'art d'un museu i es decidís fer-ho en funció de l'estil artístic (classificació per corrents com el barroc, l'impressionisme, o el surrealisme) o si es preferís fer-ho a partir de la cronologia (obres segons l'època en què van ser pintades). Totes dues són formes de classificar les obres d'art on el que varia són les relacions que s'estableixen entre les pintures, però no les obres en si que formen la col·lecció del museu. Amb aquest exercici de categorització, el monisme neutral permet resoldre el problema de la relació entre ment i matèria i escapar del reduccionisme que vol definir les característiques de la ment en termes exclusivament físics i funcionals. La qual cosa, però, no evita que hagi d'explicar què és aquesta substància —categoria o entitat— neutral subjacent i com es relaciona amb la ment i la matèria.

Des del monisme neutral és tan lícit dir que la ment és el cervell com que el cervell és la ment, perquè la substància neutral es pot manifestar de les dues formes. Al no poder tenir accés a la substància neutral, només es pot conèixer la realitat en tant que adopta el caràcter mental o material, d'altra forma resta oculta. Plantejat així, tenim que existeix una única realitat —neutral— que no és ni mental ni material. Senzillament, s'assumeix que la realitat és material quan se'ns apareix a través del sistema sensitiu, i que és mental quan ho fa a través de la introspecció. Això vol dir que la consciència pot ser entesa com un fenomen mental i, alhora, material, donat que la relació de la subjectivitat amb la manifestació física dels processos cerebrals s'explica identificant la subjectivitat amb la dependència de la manifestació específica de cada cervell individual (Nagel, 2002). De realitat subjacent només n'hi ha una, de la qual podem captar diferents aspectes. Però, la diferència entre els aspectes que captem a través de processos mentals i materials revela dues formes que té la realitat de manifestar-se, no les

diferències en la mateixa realitat o l'existència de realitats independents —com sí que fa el dualisme, per exemple—. Per tant, la diferència que hi ha entre les manifestacions mentals i materials de la substància neutral no és ontològica, sinó epistemològica.

La identificació de la substància neutral, segons el que hem vist, cau sota dos criteris o restriccions que sabem que ha de complir. El primer és que no ha de ser una entitat exclusivament material ni tampoc exclusivament mental. I el segon, que ha de ser un element irreductible a l'hora d'entendre els conceptes de ment i matèria, és a dir, que a partir d'aquest element es pugui explicar tant la seva manifestació mental com la material. Al respecte d'això, Mach considera que concebre com a única la substància material no permet satisfer aquest segon criteri, perquè «no podem entreveure cap possibilitat de representar qualsevol experiència mental en termes dels elements habitualment emprats en la física» (Mach, 1976, p. 12). Altres han suggerit com a criteri de neutralitat la possibilitat de poder identificar un element, ja sigui a través de les lleis de la física com de la psicologia, siguin aquestes les que siguin (Silverberg, 2003). Davant tal varietat de criteris i les dificultats de definir un element que no podem conèixer —només la seva manifestació material i mental—, han estat molts els candidats que s'han disputat el títol de substància neutral: els *quàlia*, l'espai-temps, o la noció de relació.[42]

Malgrat oferir una solució al problema de la relació entre la ment i el cos —simplificant-lo a una sort epistèmica que trasllada la qüestió a una realitat incognoscible—, les dificultats per a definir què és la *substància subjacent* és a la base de la majoria de crítiques que es llancen contra el monisme neutral. Que matèria i ment siguin les manifestacions de la substància neutral no explica satisfactòriament com les propietats mentals poden causar efectes sobre les propietats materials de la realitat i viceversa, ja que no dona una raó de com les intencions, les creences i altres fenòmens mentals influeixen en el comportament físic dels individus (Papineau, 2002).

Altres crítiques assenyalen una qüestió fonamental: com es pot saber si la ment i la matèria són veritablement la manifestació d'una única substància subjacent, i no de vàries. Aquesta objecció també ataca la pròpia naturalesa monista del neutralisme, entenent que no és capaç de resoldre el dualisme. De l'anterior es deriva que el monisme neutral sembla resoldre —només aparentment— el dualisme quan postula una realitat subjacent que es pot organitzar mental o materialment, però, realment, no ho fa, donat que segueix existint una separació irreconciliable

[42] Russell considera que la noció que més s'escau per definir la naturalesa intrínseca de les coses és la de relació que, al ser la més dèbil i general possible, permet situar-la en un pla que no és ni mental ni material (B. Russell, 1992, p. 80-81).

entre la perspectiva mental i la material (Metzinger, 2003), fins i tot podem arribar a veure que manté una concepció triàdica de la realitat semblant a l'estil cartesià: ment (*res cogitans*), matèria (*res extensa*) i substància neutral (Déu).

La substància neutral i el déu spinozista guarden certes semblances, en tant que única substància existent, on ment i cos són els atributs de déu que l'ésser humà pot conèixer, o el noümen kantià que defineix la cosa en si, l'existència pura i lliure de qualsevol representació. En referència això, Spinoza considerava que cada entitat particular és una modificació de la substància divina, la qual es manifesta a si mateixa a través de la seva *extensió* i *pensament*. El primer atribut fa referència a la seva existència corpòria en l'espai i el temps, i el segon vindria a representar la ment de les persones i els animals. De tota manera, també sostenia que les entitats inanimades deuen la seva existència al fet que formen part del pensament diví, deixant veure una animació universal en la qual totes les entitat existents–objectes, plantes o animals–són animades. Això suposa una sortida del problema ment-cos, en tant que formula un monisme segons el qual la realitat és la manifestació de la substància divina (Damasio, 2009, p. 178-182). El que hauríem d'examinar és fins a quin punt podem acceptar que no es caigui altre cop en un dualisme, al considerar que les dues manifestacions de la substància divina, tot i formar part de la mateixa essència, siguin dues manifestacions diferents.

També guarda similituds amb el pluralisme epistemològic en entendre que hi ha diverses formes de tenir accés a la realitat. Però, el neutralisme estableix que les manifestacions de la realitat són la ment i el cos, mentre que el pluralisme no entra a quantificar-les; ni subscriuria que la realitat estigui oculta i que només la puguem conèixer en aparença, sinó que defensa que la coneixem de forma parcial.

En qualsevol cas, el monisme neutral no aconsegueix oferir una identificació rigorosa i consistent de la consciència, donat que la seva concepció d'aquesta oscil·la entre una manifestació mental i material de la substància subjacent, on es pot adoptar una explicació idealista o materialista segons convingui.

En suma, el monisme neutral proposa una sortida del problema ment-cos que poc interessa a la ciència i al camp de la IA, perquè la neutralitat de la substància subjacent impedeix crear un model de la realitat, només un model de la manifestació mental o material de la realitat. No es pot fer un model d'allò que és incognoscible per definició. Cal assenyalar, no obstant, que no poder modelitzar la realitat en si mateixa no impedeix conèixer-la, i que fer un model de la ment humana, en aquest context, significaria comprendre la manifestació mental i material de la

realitat, encara que no es pugui conèixer essencialment. Només cal veure que tots els models que s'han fet sobre la ment poden revelar únicament característiques funcionals de la consciència, quedant inaccessibles altres aspectes de la subjectivitat (intencionalitat, creences consciència). El problema del neutralisme per a la ciència és que no aporta coneixement sobre com és que una entitat —la seva manifestació material— pot arribar a ser conscient de si mateixa, ni molt menys de si una màquina podria arribar a ser conscient. No permet conceptualitzar la realitat a través dels termes mesurables i reproduïbles que necessita la ciència per a comprendre-la.

La concepció immaterialista nega la manifestació material de la realitat, cosa que motiva encara més preguntes de les que pot contestar, perquè semblaria que atempta contra el coneixement científic que tenim i l'idea d'una realitat tan tangible com ens diuen els sentits. El monisme neutral, per altra banda, pot conciliar ment i cos, però el preu que paga és una substància incognoscible, que no resulta gens senzill identificar i molt menys estudiar. En el seu intent per unificar la realitat a una sola substància, el monisme materialista opta per la matèria com a principi fonamental. Encara que la matèria sigui la base de tot el que existeix, hi ha aspectes de l'existència que no es poden reduir a una amalgama de components materials, com poden ser els processos o les propietats que tenen les diverses configuracions de la matèria. Això no evita crítiques sobre com s'ha d'organitzar la matèria per desencadenar aquests processos o posseir aquestes propietats i, de fet, explicar-ho és un dels principals reptes amb què s'enfronta el monisme materialista.

Hi ha moltes formes d'entendre el terme materialisme, donada la seva naturalesa polisèmica, motiu pel qual convé aturar-nos a veure en quin sentit l'entendrem aquí. El materialisme filosòfic defensa que tot el que existeix és material i comparteix gran part de la seva forma d'entendre la realitat amb el *naturalisme*,[43] pel fet que els dos corrents insisteixen en que l'univers està format exclusivament d'objectes concrets. També es caracteritza per rebutjar l'existència d'éssers vius incorporis. És en relació a com consideren les característiques de la matèria on aquests dos enfocaments deixen de coincidir. Pels naturalistes la matèria és allò que la ciència (la física, la química, la biologia, etc.) investiga. En conseqüència el naturalisme no pot acceptar l'existència d'altres classes de matèria com la matèria pensant, la matèria social, la matèria artificial o la matèria semiòtica (Bunge, 2015, p. 222).

[43] Doctrina que nega l'existència de res fora de la naturalesa, com ara entitats espirituals, divines o qualsevol tipus de sobrenaturalisme (Bechtel, 2007).

El mecanicisme considera l'univers com un conjunt de cossos, de forma que la mecànica satisfà les condicions de suficiència i necessitat per a la comprensió de tot el que en l'univers succeeix. Un plantejament com aquest desafia moltes de les intuïcions que tenim de nosaltres mateixos, ja que conclou que el funcionament dels animals i de les persones no és diferent al d'una màquina. Juntament al materialisme clàssic, i fruit de la dialèctica hegeliana, sorgeix una nova forma de materialisme a mans de Marx i Engels: el materialisme dialèctic. Recuperant la noció de contradicció de la lògica de Hegel, d'acord amb les lleis de la dialèctica, afirma que tot el que existeix és una unió d'oposats, que tot canvi és fruit d'aquesta lluita. Paral·lelament al materialisme dialèctic, Marx i Engels també van idear el materialisme històric. Aquí es proposen fonamentar les societats sobre les necessitats biològiques i l'interès econòmic. D'aquesta manera es fa viable entendre el món de forma integral, pel que fa la naturalesa i la societat; revelant la base material de la societat, de l'economia i de les lleis que determinen el seu desenvolupament. Finalment, de la unió entre materialisme i cientificisme sorgeix el materialisme científic, que proposa la idea de que tot el que existeix en l'univers serà estudiat de forma més pertinent si es fa seguint el mètode científic.

Segons el monisme materialista el problema no rau en com ho fa la ment per establir una relació causal amb el cos; o en com pot ser que una entitat espiritual estigui connectada amb una altra de material; sinó en el fet de no distingir correctament un objecte material com és un cervell, dels processos que pot desencadenar com, per exemple, les emocions. De la mateixa forma que es poden establir cadenes causals que relacionen les nostres ments amb la *realitat exterior* o objectes que no són ments entre si (a través de las lleis físiques) —encara que ningú els estigui observant o percebent[44]—, també s'estableixen relacions causals dins els cervells. Tals relacions causals poden ser analitzades com a processos exclusivament materials, que cauen sota el control de les lleis de la bioquímica i la biofísica i en, alguns casos particulars, les propietats que desencadenen aquests processos es poden considerar mentals.

Per al monisme ha de ser possible explicar l'aparició de propietats mentals a partir de propietats no mentals o d'entitats que no tinguin propietats mentals. De no ser així, hi ha dues opcions (i una d'addicional). O bé s'hauria d'acceptar que les propietats mentals són propietats primitives de la realitat, subjacents i inexplicables mitjançant propietats més primàries o bàsiques, conduint a —com a mínim— un dualisme substancial; o bé plantejar que si les propietats mentals són irreductibles a termes materials, podria ser perquè són il·lusòries i

[44] Contràriament a la defensa idealista segons la qual les cadenes causals existeixen únicament quan hi ha un subjecte que les pugui percebre.

inexistents, un punt de vista defensat per autors com Quine (2013), Churchland (1990) i Rosenberg (1981) en el marc del *materialisme eliminatiu*.[45] Una opció addicional seria considerar que la matèria pot donar lloc a diferents processos o propietats segons les disposicions que adopti, de forma que es pot distingir la matèria inerta, la matèria viva i la matèria mental segons els processos que pot desencadenar, i que aquests processos emergeixin depèn de la manera en què s'organitzi la matèria. El concepte emergència, encunyat per George Lewes[46] el 1877, és la base de l'emergentisme, la concepció ontològica que adopta aquesta tercera opció.

D'acord amb l'emergentisme, determinades propietats de certes entitats no es poden explicar o predir a partir de les seves parts constitutives, i aquestes propietats es diu que són emergents. Una propietat emergent d'un objecte és una propietat —qualitativament nova— que no posseeix cap part de l'objecte ni cap component d'aquest per separat. En paraules de Searle (2000, p. 30) «no és una propietat de cap dels elements individuals [d'un sistema], i no es pot explicar com un agregat de les propietats d'aquests elements». L'emergència és una propietat que fa referència a propietats de segon ordre, en la mesura en que caracteritzaria les propietats de primer ordre com a «emergents», per distingir-les d'aquelles que no ho són (O'Connor, 1994). Així, les propietats emergents es distingeixen de les que no ho són en el sentit que no poden ser deduïdes o explicades només a través del coneixement dels seus components de forma individual; mentre que les propietats no emergents són aquelles que es poden atribuir als components individuals d'un sistema.

4.3 Materialisme i neurociència

El coneixement sobre el funcionament del cervell i el sistema nerviós ha motivat que les neurociències se subscrivissin —de forma convenient— a l'ontologia materialista, al considerar que s'ha pogut associar l'activació de parts del cervell amb la conducta, trobant el correlat físic de l'activitat mental.[47] Fer un mapa de la ment en el cervell és l'objectiu subjacent de les neurociències. Al capdavall, la finalitat de la neurociència és explicar els processos mentals i el

[45] És el punt de vista que defensa que els nostres estats mentals mostren un conjunt d'il·lusions perquè realment només poden existir condicions i processos materials, així les teories que es basen en el sentit comú són essencialment defectuoses i han de ser substituïdes per una *neurociència completa* (P. M. Churchland, 1981).

[46] Tot i que va ser John Stuart Mill —professor de Lewes— l'impulsor del terme emergència, fent notar que la síntesi de l'aigua a partir de l'hidrogen i l'oxigen involucra l'emergència de certes propietats, característiques de l'aigua, que els manquen a ambdós elements per separat (Blitz, 1992).

[47] Seria d'interès estudiar si la concepció materialista de la ment a la qual s'adhereixen les neurociències és fruit de les proves científiques que així ho corroboren, o si és la concepció materialista la que ha esbiaixat els aparells tecnològics per només corroborar l'existència d'allò material.

comportament en termes de l'estructura i la funció de les regions corresponents del cervell i la resta del sistema nerviós (Purves et al., 2008, p.57). La idea de *mapejar* la ment és la intenció que té el *Human Brain Project*, que des del 2013 treballa per avançar en el camp de la neurociència, la informàtica i la medecina relacionada amb el cervell.[48] Té com a finalitat resumir el coneixement que fins ara es té del cervell i simular el seu funcionament en sistemes d'IA.

El projecte, finançat per la Comissió Europea amb més de mil milions d'euros, també s'anuncia com un dels pioners en proporcionar cervells artificials gràcies a la neurociència artificial basada en models informàtics, que eliminaria els problemes ètics de la recerca amb cervells reals. Creuen que a través de la neurociència predictiva, que utilitza l'aplicació dels sistemes d'aprenentatge profund (*deep learning*) més sofisticats, és possible predir com es connectaran els diversos conjunts de neurones i obtenir un mapa complet de les connexions neuronals del cervell, un mapa que rep el nom de *connectoma*. Aplicant aquest mètode, diuen haver aconseguit un 90% d'èxit en les prediccions. Cal marcar, però, que les prediccions es refereixen a cervells de *peix zebra*[49] (Vladimirov et al., 2018).[50] Encara que actualment es disposi del mapa del cervell humà més detallat que mai s'ha tingut (Zachlod et al., 2023), s'hauria d'esperar que una tecnologia destinada a fer un mapa del cervell humà i predir com s'esdevindran les seves connexions neuronals gaudís d'un error molt menor al 10% en la predicció de les connexions neuronals del peix zebra.

A mitjans del segle XIX, ja es creia en la hipòtesi que a través de la neurolingüística era possible establir una relació unívoca entre les regions del cervell i les funcions mentals, que confirmaria un *localitzacionisme* cervell-ment. Nombrosos estudis han constatat que una lesió en un costat del cervell suposava una pèrdua de control del moviment del costat oposat del cos (Pia et al., 2004); que depenent de la regió del cervell que estigui afectada, una afàsia pot ser semàntica o sintàctica (Geschwind, 1970); que certes lesions provoquen la pèrdua de paraules que s'utilitzen per referir objectes inanimats, però no objectes animats —o viceversa— (Lowder i Gordon, 2015); que les persones bilingües amb danys cerebrals poden perdre competència en un dels idiomes depenent de la zona afectada (Paradis, 2004).

[48] Així és com queda definit en el seu web: www.humanbrainproject.eu.

[49] Les larves del peix zebra, al ser transparents, permeten seguir el moviment de les seves cèl·lules, l'activitat dels seus gens o de les seves neurones en temps real, fet que el converteix en un model ideal per a conèixer millor la biologia humana.

[50] El cervell humà conté 100.000 milions de neurones, mentre que el del peix zebra unes 100.000.

Altres investigacions sobre lesions cerebrals, han suggerit que aprendre pot ser una funció de diversos sistemes del cervell, però que la majoria de vegades recau en un circuit neural concret que és necessari i suficient per a l'aprenentatge (Thompson, 2005). Sigui dit de pas, aquesta descoberta indicaria que el coneixement es pot localitzar fàcilment en el cervell. Amb l'aparició de les tècniques de neuroimatge, la neurociència cognitiva va semblar revelar aquesta correlació entre les àrees del cervell i els estats mentals, donant més versemblança a la tesi del localitzacionisme materialista. En l'*àrea visual primària* és on es localitza la visió; l'olfacte en el *bulb olfactiu*; l'*amígdala* es veu involucrada en la funció de tenir por; l'*ínsula*, el *còrtex prefrontal* i el *còrtex cingulat* s'activen davant l'exclusió social; avaluem i prenem decisions amb els *lòbuls frontals*; etcètera.

No obstant això, aquestes associacions entre zones del cervell i les funcions mentals que activen no confirma el pressupòsit localitzacionista. Altres investigacions informen de que una regió del cervell pot tenir diverses funcions i formar part de diversos circuits neuronals. Així com la visió activa el còrtex visual primari, quan es mira amb atenció i deteniment també hi participen l'àrea frontal i parietal (Bressler et al., 2008). I de forma similar, ensumar (*a consciència*) a banda del bulb olfactiu, també requereix del lòbul temporal (Sobel et al., 1998). És a través de la conjunció d'aquestes àrees per a realitzar noves funcions, que podem explicar la diferència entre veure i mirar, olorar i ensumar o sentir i escoltar. A banda dels processos especialitzats de cada regió, el cervell també compta amb un sistema neuronal que estableix connexions de llarga distància que interconnecta gran varietat d'àrees cerebrals de forma coordinada i variable (Dehaene i Naccache, 2001); i s'ha pogut contrastar aquests tipus de connexions distants, mitjançant l'aplicació d'*estimulació magnètica transcraneal* en una zona concreta del cervell, i observant com s'estén a regions distants (Bestmann et al., 2004).

Si se suposa que els cervells han estat *dissenyats* de tal forma que cada part executa una única funció, és normal que aquest enfocament sembli contraintuïtiu, caòtic i poc eficient, davant la necessitat de cooperació entre àrees tan allunyades. Però estem justificats a creure que el dissenya d'aquest aparell ha estat a mans d'una evolució que produeix sense intencionalitat, sense fixar-se en altre criteri que l'adaptabilitat, deixant en el camí pretèrits defectes. Per això, no es pot establir una relació una-a-una de les regions del cervell per a cada funció, donat que hi ha moltes funcions que es donen combinant diverses àrees.

La neurociència hereta la perspectiva emergentista del monisme materialista per poder descriure tot el que experimentem com a ment en termes de processos cerebrals. Entén que el cervell és el substrat fisicalista de la ment, d'igual forma que les màquines computacionals ho

són de la IA. Defensa que la ment emergeix de la matèria i que és a través de la anàlisis del cervell que la podrem explicar i entendre. Si la neurociència, que investiga el cervell humà, el sistema nerviós central i el seu funcionament, aconseguís identificar i descriure cada procés que té lloc en el cervell —suposant que això fos concebible—, s'hauria d'acceptar que també desxifraria les incògnites sobre què són la ment i la consciència? Certament, sense cervell i sistema nerviós no hi hauria ments. Però és desencertat d'aquí extreure'n que coneixerem en el mateix detall la ment amb el coneixement que tinguem del cervell, perquè una condició necessària dista de ser una condició suficient. Creure que podrem entendre de forma completa la ment tan bon punt comprenguem el funcionament del cervell, és tan il·lusori com afirmar que s'entén per complet un quadre simplement perquè es comprèn la química dels pigments i la física de la llum. En conclusió, fer un model del cervell no equival a fer un model de la ment, i encara que es fes un mapa de la ment en el cervell, no seria una còpia d'aquesta, perquè el model resultant podria tenir característiques completament diferents del que es pretén comprendre i explicar amb aquest.

4.4 El model computacional de la ment: el funcionalisme

El funcionalisme se situa com a eix central en la filosofia de la ment contemporània, i en part això es deu al fet que és una concepció que ha permès una fonamentació ontològica i epistemològica de les troballes en psicologia, ciència cognitiva i neurociència (Van Gulick, 2009), provocant que els nous enfocaments sobre el problema de la ment hagin de posicionar-se, si més no, a favor o en contra.

Per eludir el reduccionisme cervell-ment i oferir una altra alternativa monista materialista al problema ontològic de la ment, el funcionalisme escapa de la identificació dels estats mentals amb l'activació neuronal, per centrar-se en l'aspecte funcional de la ment, és a dir, en les funcions que efectua. La idea és que les propietats mentals són propietats funcionals, i entenem que una entitat té una propietat funcional quan té la capacitat de causar un efecte en un sistema. Tenir una funció —o desencadenar un paper causal— significa establir relacions causals amb estímuls pròxims (*input*), respostes pròximes (*output*) i altres estats (Fodor, 1985). Podem destacar dos aspectes de la relació que guarden les funcions amb la seva materialització. Per un costat, una funció és independent de la seva realització material. Per exemple, una columna pot estar feta de formigó, d'acer o de fusta i complir la mateixa funció. I per l'altre, la funció també pot ser pensada en abstracte i sense matèria perquè, el concepte de *columna* no és una columna. La classificació de les propietats funcionals és més abstracta que la de les propietats materials, de manera que la classificació de les propietats funcionals i la de les propietats materials no tindrien

perquè mantenir una rigorosa relació una-a-una. En conseqüència, les propietats funcionals no poden ser idèntiques a les propietats materials. Seguint amb l'exemple, una columna no consisteix a tenir unes determinades propietats materials, tot i que tota columna particular és un objecte material particular.

És raonable preguntar fins a quin punt aquesta perspectiva està vinculada al materialisme. Al cap i a la fi, semblaria que el funcionalisme desvincula les propietats materials de les propietats funcionals al no necessitar que s'estableixi una correspondència directa entre elles per explicar-les. El cert és que la relació que s'estableix entre les propietats materials i funcionals és de caràcter asimètric (unidireccional), perquè poden existir objectes tals que les seves propietats materials difereixin i no ho facin les seves propietats funcionals, però no a la inversa: no hi ha una diferència funcional sense una diferència material. No poden donar-se casos en què objectes materialment idèntics difereixin en les seves propietats funcionals, si són materialment idèntics també són funcionalment idèntics.[51] S'estableix una relació de *superveniència* que —per mantenir-se dins el materialisme— determina la necessitat d'un substrat material per a la funció i no viceversa, definint la impossibilitat que un objecte canviï en algun aspecte funcional sense canviar en algun aspecte material (Davidson, 1980, p. 214). La relació de dependència metafísica entre matèria i funció radica en el fet que les estructures funcionals es poden realitzar en diverses configuracions materials, per tant, que les disposicions concretes de la matèria són una condició suficient per a determinades funcions, però no una condició necessària. Multitud de configuracions materials poden realitzar la funció de ser una columna, on cada configuració és suficient per a realitzar aquesta funció, però no necessària per tal que un objecte sigui una columna. Aquesta característica segons la qual una mateixa funció pugui ser implementada per configuracions materials diverses es denomina *realitzabilitat múltiple*.

L'enfocament funcionalista posa en escena que les ments no deixen de ser organitzacions de propietats funcionals (funcions) que poden realitzar conjunts materials determinats. En virtut d'això, podem considerar aquest enfocament materialista com un que no necessita associar cada estat mental amb una correlació material específica. En lloc d'això, es basa en una relació de supervinència, on es requereix la matèria per a la realització dels estats mentals, però aquests poden ser implementats per diverses configuracions materials (realitzabilitat múltiple). En efecte, significa que la consciència és realitzable per múltiples sistemes, i que pot existir en una pluralitat de substrats que no siguin necessàriament cervells biològics. La importància que el

[51] Una característica que queda definida per la Llei de Leibniz: dos objectes són idèntics si i només si tenen exactament les mateixes propietats.

funcionalisme posa en la forma en què s'organitzen les propietats funcionals —en detriment de la configuració material—, indica que qualsevol sistema que tingui una organització adequada de les seves propietats funcionals pot presentar estats mentals, independentment de les propietats materials d'aquest.

El fet que la configuració material sigui una condició suficient —i, molt important, no necessària— per a la col·lecció funcional a la qual anomenem consciència, pot fer pensar que un sistema operatiu amb una estructura funcionalment similar a la ment humana també tindria consciència. No existiria, així, una limitació material perquè l'ordinador amb què s'estan escrivint aquestes paraules sigui conscient, només limitacions funcionals. Enfront del que podria semblar una debilitat per a la teoria, els seus adeptes consideren que aquesta característica del funcionalisme és una virtut, donat que confereix molta flexibilitat alhora d'atribuir consciència a tota classe d'entitats (Heil, 2000, p. 89). Consideren que negar que altres sistemes funcionalment estructurats puguin tenir consciència és un prejudici antropocèntric. Contra aquest prejudici és contra el qual el funcionalisme vol lluitar, atès que no seria absoluta la impossibilitat d'una consciència artificial amb una distribució material de silici (Kurzweil, 2019, p. 202).

Pel funcionalisme, encara que els estats mentals siguin propietats funcionals i les propietats funcionals sobrevinguin a les propietats materials, els estats mentals no són propietats materials. Si bé, no podem dir que caigui en un reduccionisme ment-cervell, que redueixi la ment a processos materials neurocientificament observables, sí que redueix la concepció de la ment a un seguit de funcions. Tota la realitat mental que experimentem, assegura el funcionalisme, pot ser explicada en termes funcionals, de la mateixa manera que la física descriu el moviment d'un pèndul.

És il·lustrador l'exemple del dolor que posa (Block, 1978). De la mateixa manera que el fisicalisme diu que el dolor és un estat físic, el funcionalisme diu que el dolor és un estat funcional. El correlat funcional del dolor es defineix en termes d'entrada i sortida i no en termes mentalistes. L'exemple ens proposa suposar que un dolor és causat per algun dany a la pell que causa preocupació i l'exclamació «Au!», i aquesta preocupació provoca que s'arrugui el front. Aleshores podem descriure el dolor com la propietat d'estar en un estat (funcional) que és causat per un dany a la pell que causa exclamar «Au!»; i un altre estat que causa arrugar el front. Destaca el fet que els conceptes mentals «dolor» i «preocupació» han desaparegut de la definició

funcional, i han estat substituïts per variables,[52] mentre que l'entrada i la sortida continuen presents. Perquè la definició sigui correcta, tota entitat que exclami «Au!» i arrugui el front davant un dany en la pell, estarà sentint dolor i viceversa. Fàcilment es pot imaginar un artefacte dissenyat amb pell artificial, front, una gravació que exclami «Au!» i dos estats que compleixin les relacions causals descrites, però que no senti dolor. Això últim és el que el funcionalisme negaria, al considerar que, per cada estat, mental hi ha una descripció funcional —on es tradueix tot el vocabulari mentalista en variables— que no permet que una entitat que satisfaci la descripció no posseeixi l'estat mental en qüestió.

Putnam analitza el problema des de la premissa que els estats mentals són funcions independents de la seva realització material. A través del model computacional de *màquina de Turing*, suggereix que un computador és un model de la ment i de la seva relació amb el cos. No és d'estranyar que la idea de màquina de Turing fos clau per concebre la IA, ja que el funcionalisme assumeix que la realització física d'una màquina de Turing és un computador, i que computadores físicament diferents poden ser realitzacions de la mateixa màquina sense deixar de ser funcionalment equivalents. En aquesta ocasió, Putnam considera la màquina de Turing com un *autòmat probabilístic*[53] i a explica que un estat mental, com ara sentir dolor, és un estat funcional de l'organisme que sent dolor (Putnam, 1975).

Tenir estats mentals, afirma Putnam, consisteix en disposar d'una organització funcional adequada, i tenir un estat mental particular, com el dolor, és trobar-se en un estat funcional particular. Així, la relació entre la ment i el cos és anàloga a la relació entre el suport lògic o software i el suport físic o hardware d'un ordinador. L'activitat mental dels organismes hauria de ser entesa des d'un punt de vista computacional, com un conjunt d'operacions logicoformals expressables amb símbols i representacions (Fodor, 1984, p.11), fet que revela el vincle del funcionalisme amb una concepció representacional de la ment. Per aquest motiu, s'ha denominat *funcionalisme computacional* als enfocaments que han concebut el funcionalisme d'aquesta forma, establint que els problemes filosòfics derivats de la relació entre ment i cos queden resolts en entendre aquesta relació com la que hi ha entre els estats lògics i els estats estructurals d'una màquina.

---

[52] L'explicació funcional és que els estats mentals poden ser sotmesos a una anàlisi causal dels seus conceptes, de forma que conceptes com «dolor» o «preocupació» serien estats que tenen la propietat de ser causes de certs efectes (*input*) i ser efectes de certes causes (*output*) (Armstrong, 2002).

[53] En comptes de determinista, com estableix el model de Turing, que vol estudiar la qüestió plantejada pel matemàtic alemany David Hilbert sobre si les matemàtiques són decidibles.

Si la implementació material d'un programa (software o ment) és múltiple i indeterminada, es revela que no existeix un problema ment-cos, perquè ningú es plantejaria la relació entre software i hardware com un inextricable problema metafísic.

4.5 Consciència artificial i funcionalisme computacional

Ni tan sols hi ha un consens científic sobre què és la consciència, tanmateix reproduir-la de forma artificial a través de sistemes d'IA és una de les empreses més ambiciosa del sector tecnològic i una de les notícies més recurrents als mitjans de comunicació. Sembla que tothom està familiaritzat amb la consciència, però ningú sap exactament què és. D'entrada, quan ens referim a la consciència ho podem fer generalment de tres formes diferents: com la capacitat de jutjar la moralitat dels actes; com l'estat d'una persona que està desperta i pot raonar; i com el coneixement immediat i directe que la persona té de la seva existència. Què pot voler dir que una màquina sigui conscient? Per bé que l'accepció que es fixa en la capacitat d'emetre judicis és molt suggerent pel sector tecnològic i necessita ser examinada, quan es parla de consciència artificial es fa referència generalment a les propietats que ha de tenir un sistema d'IA per dir que té coneixement de la seva pròpia existència subjectiva.

Quan parlem de consciència, sempre ho fem de la consciència d'algú (un subjecte), el que es coneix com la *condició de propietat de la consciència*. Els problemes que es deriven d'intentar explicar la consciència i, en especial, la d'altres éssers, és del que Nagel (Nagel, 1974) alerta quan es pregunta sobre com percep el món un ratpenat i com podem conèixer el grau de consciència de la realitat que té. El cas, explica, és que no podem tenir accés a aquesta informació perquè no entenem el món mitjançant els mecanismes que fa servir un ratpenat, en una línia similar a la de Wittgenstein quan diu «si un lleó pogués parlar, no el podríem entendríem» (Wittgenstein, 2012a, p. 511). I no només no podem saber com és la imatge del món que es crea un ratpenat o un lleó, sinó que tampoc podem saber com és la imatge del món que té qualsevol altra persona, només el que ens expliqui que experimenta la persona en qüestió, fet que coneixem com la *privacitat de la consciència*.[54] Aquest és un dels motius pels quals les ciències empíriques fracassen en explicar la consciència, ja que una cosa és l'experiència que suposa la consciència i una altra és com informa un subjecte —de consciència— d'aquesta experiència.

[54] Al respecte de la privacitat de la consciència, es pot objectar que si mai no poguéssim saber la imatge del món que té una altra persona i la consciència només fos d'accés privat, només podríem fer hipòtesis molt generals sobre el món, i tindríem grans dificultats per arribar a acords sobre com és la realitat.

De la consciència també es deriven altres problemàtiques, com determinar en quin moment en una persona es desperta la consciència. Neix amb la capacitat de ser conscient o aquesta apareix en el nadó a mesura que va interactuant amb la resta d'humans? Hi ha un acord generalitzat segons el qual és amb la capacitat de reconèixer la seva imatge en un mirall, a partir dels dos anys aproximadament, quan un nadó exhibeix consciència de si mateix (Rochat, 2003), tot i que un sector és reticent a acceptar que el propi reconeixement sigui un factor suficientment exhaustiu de consciència (Povinelli, 2001).

En general, el gran problema d'una aproximació de la consciència des de la ciència és que, si resulta ser una capacitat íntima que els éssers tenen i no poden transmetre, com l'avaluem i accedim científicament a ella? Es podria determinar a través d'un examen conductual, observant el comportament de les entitats per a determinar si són conscients o no (Irvine, 2013). Però això porta el problema afegit de la imitació, és a dir, que una entitat tal com una IA podria imitar tan bé el comportament d'un ésser conscient que se la donés per conscient, quan el que estaria fent internament serien processos exclusivament mecànics que no tenen res a veure amb la consciència.[55]

Amb la finalitat d'establir si una IA podria arribar a ser conscient, un grup d'experts ha presentat un document (Butlin et al., 2023) on adopten el mètode del funcionalisme computacional per a estudiar la validació empírica de les condicions per a la consciència. Els autors defensen que les teories neurocientífiques de la consciència gaudeixen de reconeixement empíric significatiu i que poden ser útils per a analitzar la consciència de la IA. Consideren que, a través d'aquestes teories, és possible identificar funcions que són necessàries i suficients per a la consciència, i que des del funcionalisme computacional funcions similars són suficients per dir que un sistema d'IA és conscient.

El funcionalisme computacional parteix de dos supòsits: el funcionalisme i el computacionalisme. El supòsit funcionalista es fixa en el tipus de funcions que un sistema pot fer i, depenent de les funcions que pugui realitzar, es dirà que té unes capacitats o altres: si es pot determinar quines funcions fan que un ésser sigui conscient, també es pot examinar si un sistema té les mateixes propietats funcionals.[56] El supòsit computacionalista assumeix que

[55] De forma similar a com es presenta el Test de Turing, on s'efectua un examen conductual, com veurem més endavant.

[56] Entenent «propietat funcional» com aquelles que un objecte posseeix en virtut de la seva aptitud per a desenvolupar un paper causal en un context determinat (Moya, 2006, p.78).

partint de la idea que aquestes funcions es refereixen a processament d'informació, manté que la consciència és una forma de processar la informació. Per tant, considera que els sistemes d'IA que processen grans quantitats d'informació són bons candidats per aplicar aquest mètode.

La forma en què aquest mètode analitza si la IA pot ser conscient consisteix a examinar l'arquitectura interna dels sistemes i els processaments que realitzen. Llavors, es mira si aquests processos encaixen amb els processos que tenen lloc en la ment humana, i que algunes teories neurocientífiques han identificat com les funcions que s'han de donar per a la consciència. Basant-se en aquestes teories neurocientífiques sobre les funcions de la consciència, en el mencionat estudi (Butlin et al., 2023) s'elabora una llista d'ítems que indicarien que un sistema d'IA seria conscient.

Aquí comencen els problemes, donat que la gran dificultat està en que no hi ha un consens entre la comunitat neurocientífica sobre què és la consciència, encara que existeixin multitud de teories que l'intentin explicar. Per això, els autors del document (Butlin et al., 2023) expliquen que el compliment d'aquest llistat d'ítems per un sistema d'IA no garanteix que sigui conscient, tan sols ho fa més probable quants més ítems sigui capaç de complir. Així doncs, el que presenten són criteris de possible consciència, en cap cas es pot determinar si un sistema és conscient o no a través d'aquest mètode.

Que s'adopti el funcionalisme computacional ens revela que els autors se subscriuen a una perspectiva del materialisme monista, perquè entenen que alguns estats cerebrals són experiències conscients i altres no, i que la tasca de la neurociència és entendre en què es distingeixen. Ara bé, no és una doctrina senzillament reduccionista, com altres que s'han vist, donat que no equipara, sense més ni més, tot estat mental a una activació neuronal corresponent, sinó a les seves propietats funcionals. La consciència sorgeix de fenòmens materials —com les connexions neuronals o els processos bioquímics—, però amb independència de l'estructura material.

Els autors (Butlin et al., 2023) es basen en diverses concepcions sobre la consciència a partir de destacats estudis i teories científiques que l'han analitzat. D'aquesta manera arriben a extreure les funcions necessàries per a la consciència i elaboren els ítems que verificaran la possible consciència d'un sistema. Una d'aquestes teories és la del processament recurrent (*Recurrent Processing Theory*), que es basa en experiments neurocientífics on s'ha estudiat exclusivament la visió, i que busca explicar la distinció entre els estats en els quals els estímuls es veuen de forma conscient dels que només estan representant l'activitat del sistema visual de

forma inconscient (Lamme, 2020). La neurociència de la visió ha observat que les neurones encarregades de processar els impulsos visuals tenen tendència a rebre una entrada (*input*) de la seva pròpia sortida (*output*). Més concretament, aquestes neurones (per exemple, X) processen informació que envien a altres neurones (diguem-ne, Y) que també processen aquesta informació, i que aquesta informació processada (primer per les neurones X i després per les neurones Y) acaba tornant a les neurones inicials (a X). Per tal de processar la informació visual, els sistemes neuronals creen bucles, i per això es diu que és *processament recurrent*. Els experiments realitzats semblen indicar que l'activitat del sistema visual és suficient per definir la consciència, sempre i que es compleixin certes condicions bàsiques. Sostenen que la distinció entre estats conscients i inconscients es correspon a diferents etapes en el processament visual. Quan el processament de la informació visual passa per diverses neurones de forma recurrent, el fet que aquest senyal torni a les neurones inicials afavoreix l'aparició de l'experiència visual conscient.

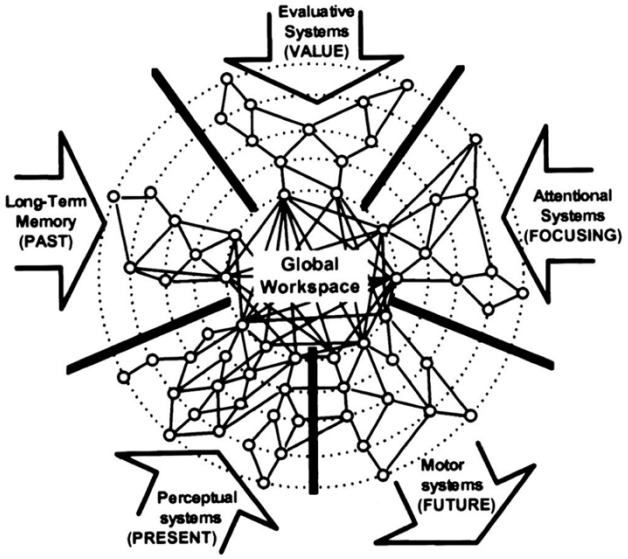
Posem el cas d'un estímul que no genera *interès* a la neurona inicial que processa la informació, el senyal que li tornarà en completar el circuit (bucle) serà molt dèbil i no es dirigirà l'atenció cap a aquest estímul —no hi haurà recurrència en el processament—. L'activitat serà suficient per a certes operacions visuals, com extreure característiques de l'escenari, però no per a l'experiència visual conscient. En canvi, quan l'estímul és prou significatiu, és produeix un processament recurrent que genera una representació conscient d'una escena organitzada, en contraposició a tots els estímuls visuals que són inconscients. Des d'aquesta teoria, l'experiència visual conscient no exigeix la participació d'àrees no visuals i, per tant, l'experiència conscient pot ser considerada com un fenomen *local*.[57] Que sigui local vol dir que, mitjançant una arquitectura recurrent, pot donar-se l'experiència conscient dins un únic mòdul, sense la necessitat d'integrar-se amb altres parts del cervell. Aplicant aquesta teoria, els autors extreuen dos primers criteris per avaluar si un sistema és conscient (Butlin et al., 2023):

a) Un sistema basat en mòduls d'entrada (*input*) que utilitzin recurrència algorítmica.
b) Els mòduls d'entrada han de generar representacions perceptives organitzades i integrades.

Una altra teoria que els autors fan servir, i que contrasta amb la del processament recurrent, és la de l'*espai de treball global* (*Global Workspace Theory*), que considera que les persones

[57] Contràriament a l'opinió d'autors que consideren que les funcions del còrtex prefrontal són necessàries per a la consciència (Malach, 2022).

tenen accés a diversos mòduls per a realitzar tasques cognitives, i que per a l'experiència conscient aquests mòduls s'han d'integrar globalment (Baars, 1988; Mashour et al., 2020) . Els mòduls als quals es té accés serien la visió —entesa com el sistema format per ulls, nervis i part del cervell on es processen les dades de la visió—, el tacte, l'olfacte, etc. Ara bé, els mòduls per si sols no són conscients. Què és el que fa que emergeixi la consciència d'aquest mòdul? Sostenen que per tal de poder treballar amb les dades de la visió, del tacte i de l'olfacte, és necessari que en la ment hi hagi un espai abstracte en el qual es representi tota aquesta informació; una sèrie de neurones que s'encarreguen de fer una representació de *tota* la informació captada. La representació ha de ser necessàriament més petita que la suma d'informació de tots els mòduls, de forma que cada subjecte ha de seleccionar la informació que representa en aquest espai, no pot representar totes les dades que li arriben (veure Il·lustració 6: representació espai de treball global (Global Workspace)).



Il·lustració 6: representació espai de treball global (Global Workspace)

Font: (Dehaene et al., 1998)

El lloc on es representa aquesta informació és l'espai de treball, i si es poden fer operacions complexes sobre aquesta informació és perquè la resta de mòduls té accés a l'espai de treball. La memòria té accés a la representació visual, el tacte, a la de la memòria i així progressivament, de forma que hi ha una espècie d'accés *global* a aquest espai de treball comú. Segons aquesta teoria, la consciència existeix per la necessitat d'integrar tots aquests mòduls i dirigir la nostra atenció cap a aquells que siguin rellevants. A partir d'aquí, en l'estudi s'afegeixen els següents quatre criteris per a la consciència (Butlin et al., 2023):

c) Els mòduls (*input-output*) han de poder treballar en paral·lel.
d) Un espai de treball amb capacitat limitada, fet que implica que hi hagi un coll d'ampolla en la informació que requereixi un mecanisme d'atenció.
e) La informació que està en l'espai de treball ha d'estar disponible per a tots els altres mòduls, amb la finalitat de poder realitzar operacions complexes.
f) L'atenció s'ha de dirigir cap a la informació més rellevant, donant lloc a la capacitat d'utilitzar l'espai de treball per seleccionar la informació de cada mòdul segons convingui i realitzar tasques complexes.

Basant-nos en aquests sis criteris (a, b, c, d, e i f), una eina com ChatGPT—capaç de realitzar multitud de tasques amb el llenguatge natural— no es podria considerar conscient, perquè no compleix el criteri a) d'utilitzar recurrència algorítmica. Els models de llenguatge fan servir xarxes neuronals profundes amb una arquitectura que no utilitzen el processament recurrent (Vaswani et al., 2023), de forma que no es generen aquests bucles en el processament dels estímuls, considerats per a la consciència d'un sistema. Encara que aquests models de llenguatge produeixin resultats (*outputs*) que s'assemblin als que podria donar un humà, internament funcionen de forma completament diferent. Un sistema pot ser entrenat per a imitar el comportament humà tot i que els mecanismes que el portin a comportar-se així no tinguin res a veure amb la forma en què funcionen les persones. Comportar-se com si es tingués consciència no significa que se sigui conscient.

Més enllà de la recurrència algorítmica, els models de llenguatge no són un bon candidat per ser considerats conscients, perquè no són un agent que pugui actuar amb l'entorn (Butlin, 2022). Tanmateix, això dependrà de la definició d'agència a la qual ens acollim. Russell i Norvig (2010, p. 34), entenen per agent «qualsevol cosa que percep el seu entorn mitjançant sensors i que actua sobre aquest entorn a través d'activadors», suposaria haver de considerar conscient a qualsevol sistema d'IA, i també a artefactes tals com un termòstat.

Els criteris per alguns casos són poc restrictius, fent que puguem considerar alguns sistemes artificials o digitals com a candidats a tenir consciència. Però també són massa restrictius perquè se supedita la possessió de consciència a l'exhibició d'un comportament, ja que es podria arribar a considerar que no és conscient una persona que per una lesió o una malaltia degenerativa (síndrome d'enclaustrament) hagi perdut la mobilitat i el control del cos.

És important reiterar els riscs de sobredimensionar l'atribució de consciència a la IA, que es deriven molts cops d'aquesta tendència humana —explicada amb anterioritat— a

antropomorfitzar els objectes i que pot portar a atribuir característiques humanes i estats mentals a entitats que no en tenen. Alhora, aquesta predisposició a antropomorfitzar l'entorn, també genera la injustificada exigència d'haver de desenvolupar sistemes que s'assemblin a les persones. Segurament perquè, si s'assemblen a nosaltres, podrem pensar que els seus objectius també o fan perquè la idea que hi ha darrera del seu desenvolupament és que puguin fer les nostres feines.

Els riscs que es deriven d'aquesta tendència a atribuir capacitats humanes a la tecnologia tenen a veure amb biaixos d'agència i es poden donar per diversos motius (Elish, 2019; Trafton et al., 2024).Són biaixos que indueixen a esperar que els sistemes es comportaran de la mateixa forma que ho fan les persones (intencionalitat o comprensió humana), conduint a una sobreestimació de les seves capacitats o a una confiança injustificada en aquests sistemes (Madhavan i Wiegmann, 2007). Inclinacions com aquestes poden portar a delegar tasques que només poden realitzar humans a la tecnologia (com l'educació o el suport emocional); a fer més vulnerables a les persones davant la manipulació (notícies falses o *fake news*); i comportar problemes relacionats amb la salut i la soledat no desitjada (trastorns mentals o substitució del contacte humà).

D'altra banda, hi ha un sector d'experts preocupat pels riscs de subestimar l'atribució de consciència, donat que sostenen que no considerar conscient a un sistema que realment ho és, pot incórrer a faltar a les obligacions ètiques existents respecte a aquest. Consideren que amb la creixent sofisticació dels sistemes d'IA, inevitablement serà necessari plantejar-se qüestions vinculades amb el seu estatus moral (Metzinger, 2021; Schwitzgebel & Garza, 2020; Shulman & Bostrom, 2021). L'assignació d'agència ètica té a veure amb la consideració ètica que mereix una entitat i alguns autors (com veurem a l'apartat 6.3) apunten a que està relacionada amb la capacitat que tingui per a experimentar patiment (Singer, 1987).

Es pot estar d'acord al voltant que qualsevol entitat que pugui patir conscientment mereixi consideració moral. No obstant això, no està gens clara la relació entre ser conscient i l'estatus moral. Ens haurem de preguntar sobre si ser conscient és una condició suficient per a la consideració moral. Dels criteris que s'han descrit (Butlin et al., 2023), com a molt es podria arribar a l'atribució de consciència d'un sistema d'IA en funció del compliment dels criteris, tanmateix se seguiria estant lluny de poder afirmar que pel fet de que un sistema sigui conscient també pugui patir o tenir emocions. Per aquest motiu, els riscs associats amb la sobredimensió en l'atribució de consciència són els que ens hem de prendre més seriosament.

La confrontació entre les teories —de les quals es deriven els criteris que hem vistes (Butlin et al., 2023)— del processament recurrent i la de l'espai de treball global és evident, en tant que la primera afirma que l'activitat en regions cerebrals locals —concretament, en la visió— és suficient per a la consciència, mentre que la segona exigeix per a la consciència un espai de treball global on es representi tota la informació i al qual la resta de mòduls tinguin accé(Michel i Doerig, 2021). Tanmateix, això no fa que els criteris que es deriven de cada teoria per a determinar si un sistema d'IA és conscient siguin excloents. Es podria donar la consciència sota criteris locals i generals alhora o, com a mínim, que ambdues maneres de capturar la consciència estiguin íntimament vinculades, segons el que es descriu a l'estudi (Butlin et al., 2023).

Bona mostra d'això seria un experiment en el qual es presentin estímuls visuals tals com «triangle groc», «cercle vermell» i «quadrat blau» a diverses velocitats i de forma intermitent. Si un subjecte informa veure un cercle vermell i prou, direm que ha estat conscient (localment) d'haver-lo vist i, que per tant, té l'experiència conscient (globalment) d'un cercle vermell. En canvi, els altres dos estímuls dels quals no s'ha tingut consciència d'haver vist, no generaran l'experiència del triangle groc i el quadrat blau perquè no haver desencadenat un processament recurrent impedeix que se centri l'atenció en tals estímuls ni, en conseqüència, tenir l'experiència conscient d'aquests.

Per concloure la reflexió sobre l'estudi (Butlin et al., 2023), dels models basats en IA que tenim a dia d'avui, cap compliria criteris suficients per a la consciència. Si bé molts dels sistemes que s'han desenvolupat satisfan alguns dels criteris tan bàsics com operar en paral·lel —com diu el criteri c)—, o generar representacions perceptives organitzades i integrades —estipulat pel criteri b)—, no en compleixen suficients com per poder pensar que hi hagi consciència en algun d'aquests models. Encara que a través dels criteris o ítems que s'han presentat per a la consciència, trobéssim un sistema d'IA que els complís tots —una possibilitat que els autors argumenten que no quedaria tan lluny—, això no voldria dir que el model en qüestió fos conscient. Senzillament significaria que compliria tots aquests criteris i, que en base al que sabem de la consciència a partir del funcionalisme computacional —que no és l'únic enfocament per a explicar-la—, segurament estaríem a prop de reproduir la consciència de forma digital, però no d'explicar-la.

La distinció entre reproduir i explicar es pot relacionar amb la noció de competència sense comprensió de Dennett (que hem vist a l'apartat 3.2), on un sistema pot imitar comprendre o ser conscient simplement perquè és capaç de complir certs criteris o desenvolupar tasques concretes. L'èxit operatiu no implica una comprensió real o consciència, ja que aquests sistemes

actuen com caixes negres que poden replicar comportaments sense que comprenguem realment per què donen una resposta. En conseqüència, en el cas de la consciència i la IA ens hauríem de preguntar si es pot reproduir artificialment la consciència sense comprendre-la i, sobretot, amb quines finalitats.

4.6 Contra el funcionalisme

Havent explorat la capacitat per a proporcionar una comprensió fonamentada de com opera la ment des de la perspectiva funcionalista, aquí es vol analitzar en quina mesura aquest enfocament ajuda a definir les característiques exigibles a un sistema conscient, especialment en el cas concret de les tecnologies equipades amb IA. A través d'aquesta anàlisi, també es vol veure fins a quin punt podem acceptar que la consciència sigui —senzillament— una certa forma de processar la informació, només condicionada per les funcions que realitzi, a l'abast de qualsevol sistema que tingui la disposició apropiada i, per tant, que funcioni com un programa informàtic.

La virtut del funcionalisme és que no necessita vincular la consciència amb determinats processos interns del cervell (connexions neuronals, senyals elèctrics del sistema nerviós, estats neurofisiològics) o amb patrons característics de conducta. El que ens diu és que no té tanta importància de quina manera o en quin substrat es doni la funció de ser conscient, sinó que aquesta funció es realitzi. Analitzarem críticament el funcionalisme amb l'objectiu d'examinar la seva idoneïtat com a marc teòric per a la identificació de la consciència. A la llum de la recerca sobre la consciència artificial, és crucial comprendre els principis subtils del funcionalisme i avaluar la seva adequació per a aquest propòsit.

L'arrel del problema en el funcionalisme es troba justament en confondre el que hem definit com a realitzabilitat múltiple amb la tesi de la *independència del substrat*. Segons hem presentat, la realitzabilitat múltiple s'estableix com un argument en contra del reduccionisme de la teoria de la identitat (els estats mentals són estats materials), defensant que una mateixa funció pot ser realitzada per diverses configuracions materials adequades. Que el funcionalisme accepti que la consciència es pot donar en una multiplicitats de substrats (realitzabilitat múltiple), no significa que es pugui fer servir qualsevol substrat per a construir un sistema conscient (Block, 1996), o que el formatge suís pugui implementar la consciència (Michel i Lau, 2021). De manera que hi haurà configuracions materials que seran més pertinents que altres per a acomplir amb

certes funcions.[58] Per contra, la independència del substrat suposa que una funció pot ser potencialment completada per objectes que tenen bases materials completament diferents. El cas és que si el funcionalisme considera que els ordinadors o programes informàtics poden ser conscients, està confonent ambdós conceptes (realitzabilitat múltiple i independència del substrat).[59] Així, veiem que s'estaria considerant només el paper funcional de la consciència, sense posar atenció en que sigui una activitat biològica, i, en conseqüència, no s'estaria aplicant el concepte de realitzabilitat múltiple sinó el d'independència del substrat, ja que els ordinadors i programes informàtics estan fets de matèria inerta, no estan vius. Sí que podem acceptar que la consciència es realitza de forma múltiple en els humans, donat que hi ha moltes diferències entre les persones (neuronals, cerebrals, morfològiques…). Però això no significa que qualsevol realització que guardi indicis de semblança amb el que fan les persones, estigui realitzant exactament el mateix.

Un altra aspecte problemàtic del funcionalisme és deriva de no poder mantenir la identitat entre propietats funcionals i propietats mentals. Recordem breument que la tesi central del funcionalisme parteix de reconèixer que els estats mentals són estats funcionals, i aquest és un supòsit que es refereix a la identitat de propietats. Una propietat mental determinada —defensa la tesi funcionalista— ha de ser idèntica a una propietat funcional determinada, de forma que dos sistemes que es trobin en un estat funcional concret, també es trobaran en el mateix estat mental. Encara que les seves propietats materials siguin diferents, els sistemes que comparteixen organització funcional han de compartir propietats mentals, del que se segueix que la identitat funcional entre sistemes implica la identitat mental (bé per presència, bé per absència). D'acord amb això, una contundent refutació del funcionalisme és demostrar que és possible que dos sistemes funcionalment idèntics siguin mentalment diferents.

Partint d'aquesta idea, hi ha diversos crítics (Nagel, 1974; Putnam, 1988; Searle, 1980) que argumenten que no existeix correspondència entre propietats funcionals i mentals, i que hi ha aspectes inherents a la ment que són inexplicables en termes funcionalistes. Un escenari com aquest és problemàtic pel funcionalisme perquè suggereix que la disposició funcional, tot i ser necessària, pot no ser suficient per explicar els estats mentals o la consciència. Alguns d'aquests aspectes mentals que no queda clar que tinguin un correlat funcional estan relacionats amb els

[58] Els objectes que tinguin la funció de ser una columna, com hem vist, poden estar fets de multitud de materials, però hi ha límits respecte als materials que els poden configurar perquè pugui complir la funció d'aguantar l'estructura d'un edifici (no es pot fer un columna de formatge, d'aigua o de roba).

[59] Una confusió que ha estat alimentada pel físic Max Tegmark amb la idea radical de que la vida és independent del substrat (Tegmark, 2015).

estats mentals fenomenològics[60] (experiència subjectiva) i els *estats mentals intencionals*[61] (creences i desitjos). Primer ens ocuparem de l'objecció que fa referència als estats mentals fenomenològics i a continuació a la que fa referència als estats mentals intencionals.

Quant als estats mentals fenomenològics, és concebible —en contra del que diu el funcionalisme— que dos entitats comparteixin la disposició funcional, però que una tingui estats mentals i l'altra no[62]. És una qüestió de contingència: no és necessari que els estats mentals específics estiguin associats a una disposició funcional, ja que les propietats qualitatives d'aquests estats mentals no permeten una caracterització funcional exhaustiva. Potencialment, un ésser humà pot tenir una infinitat d'estats mentals, mentre que el funcionalisme no pot identificar cada un d'aquests estats amb el seu correlatiu estat funcional, donat que hi ha un nombre finit de disposicions funcionals. No hi hauria prou estats funcionals per emparellar-los amb els possibles estats mentals d'una persona.[63] Evidentment que una persona només tindrà un nombre finit d'estats mentals, però aquesta és una limitació referent a la longevitat humana, i no a una llei psicològica que restringeixi la capacitat humana per a contenir estats mentals. Veiem aquí que el funcionalisme ubica erròniament els límits de la finitud humana en la cognició i la capacitat per a tenir estats mentals, en comptes de fer-ho en l'esperança de vida.

Així, es pot entendre que la descripció funcional que podem fer d'un ésser és invariable davant els canvis d'estat mental que experimenti (McGinn, 1991, p. 196-197), fet que permet que un sistema que no senti plaer tingui una disposició funcional equivalent a la d'un sistema capaç de sentir plaer. En conseqüència, hi hauria un sistema tal que funcionalment seria indistingible d'algú que sent plaer, però sense un punt de vista subjectiu —estat mental— del que és experimentar plaer. No es podria identificar l'organització material d'aquell sistema i les seves propietats funcionals amb res que sigui ésser —subjectivament parlant— aquell sistema. Al mateix temps, cenyint-nos al funcionalisme, hauríem de dir que un sistema com l'anterior està

[60] Ens referim a les propietats qualitatives dels estats mentals i el problema de definir-les en termes de disposicions causals, sigui analítica o empíricament, i que per no desviar-nos del nostre objectiu en aquest treball no s'aborden exhaustivament.

[61] Que tenen a veure amb les propietats semàntiques de certs estats mental, propietats que han de distingir-se de les propietats sintàctiques, com veurem a continuació.

[62] Seria el cas d'éssers que no tinguessin consciència i que es comportessin com si en tinguessin, com el *zombi filosòfic* (Chalmers, 2003) o el *China brain (Block, 1978)*.

[63] Una refutació que també podem fer servir contra la idea d'una màquina conscient o el model artificial de la consciència, és que més que una màquina probabilística comptés amb memòria i temps il·limitats, només podrà computar un nombre finit d'estats de màquina, i la consciència exigeix la possibilitat d'operar amb infinits estats mentals.

sentint plaer encara que no pugui sentir absolutament res, i aquesta és una conseqüència inacceptable perquè —com hem anunciat— els sistemes que comparteixin estats funcionals també han de compartir estats mentals. Si l'experiència subjectiva és una propietat essencial de certs estats mentals, llavors no hi ha res que obligui a que sistemes mentalment diferents també s'hagin de correspondre amb sistemes funcionalment diferents. Això suggereix que les experiències conscients no són propietats funcionals. Sabem que algunes funcions dels éssers vius poden ser substituïdes per sistemes inerts (pròtesis, implants, òrgans artificials), però d'aquí no se segueix que la vida sigui idèntica a una funció. Més aviat, les propietats funcionals dels éssers vius estan subordinades a la vida i a la seva supervivència com a sistema biològic. D'igual manera, les propietats funcionals que puguin tenir els estats mentals estan vinculades a les experiències subjectives que tingui un sistema que està viu (amb tot el que això significa).

L'objecció referent als estats mentals intencionals ataca concretament la concepció del funcionalisme computacional que entén la ment com un programa de computació. L'ús competent del llenguatge implica propietats semàntiques que possibiliten la comprensió dels enunciats lingüístics, gràcies a les quals podem aprehendre el significat dels enunciats i emetre'n —intencionalment— sabent el que signifiquen. Hi ha estats mentals, com les creences o els desitjos, que tenen contingut intencional. Quan parlem d'intencionalitat (del llatí *intendere*, que vol dir tendir cap a alguna cosa), ens estem referint a que els estats mentals es dirigeixen cap a quelcom que no necessàriament ha de ser un altre estat mental; o sigui, a l'orientació que prenen els nostres estats mentals cap als seus objectes. El contingut intencional té la propietat de modelitzar la realitat de diferents formes, i de poder ser vertader o fals, propietat que té en comú amb el llenguatge. La informació que proporcionen els continguts intencionals és semàntica, perquè dir si aquesta és vertadera o falsa té a veure amb comprendre el seu significat. En aquest sentit, aquesta objecció defensa que un ordinador és una màquina sintàctica, però no semàntica. Significa que un ordinador és un instrument per a processar signes segons la seva forma, estructura i altres característiques materials, sense la capacitat de conèixer i entendre el significat que poden tenir els enunciats que processa. Aquesta idea és expressada a través de l'experiment mental de *L'habitació xinesa*, formulat pel filòsof John Searle.

John Searle és un dels precursors en tractar el tema de la IA des de la filosofia, negant categòricament que els ordinadors i computadores que puguem arribar a fabricar siguin capaços de pensar o ser intel·ligents, i que la possibilitat que desenvolupin consciència o ment és completament nul·la (Searle, 1985, p. 34). En ell trobem la distinció entre dues modalitats d'IA. Per una banda, la *IA forta*, que equipara la ment humana amb un ordinador (el mite de la ment

computadora: la ment és un ordinador i un ordinador és una ment). Per l'altra, la *IA dèbil*, que accepta certa analogia entre la ment i l'ordinador, però només per a fer tasques específiques (mecàniques, processament de grans quantitats de dades, identificació de patrons…), perquè és a partir del cervell humà que podem explicar l'ordinador, i no al revés. Searle manté que el que tenim a dia d'avui, i l'únic que tindrem en el futur, són sistemes d'IA que fan tasques molt concretes i restringides, és a dir, una IA dèbil, que per més sofisticada i excel·lent que arribi a ser en la realització de tasques, sempre haurà de ser en condicions molt delimitades.

La crítica contra la IA forta es fonamenta en els teoremes d'incompletesa de Kurt Gödel, que demostren que hi ha sistemes aritmètics que contenen enunciats vertaders, però que són indemostrables. Gödel argumenta que existeixen sistemes axiomàtics que són consistents i per als quals no es pot dissenyar una sèrie precisa de regles, sense ambigüitat i que es puguin fer en un temps finit, que en verifiqui la correcció. Això és, hi ha formules semànticament vertaderes que no es poden demostrar sintàcticament i que no serem mai capaços d'arribar a aquestes formules algorítmicament. Aquesta és una limitació que tindran sempre les màquines i la IA, que només poden funcionar a través d'una successió de passos molt ben definits. L'argument que fa servir Searle en l'habitació xinesa també parteix de la incompetència de les màquines per a operar semànticament, tot i la seva aptitud per fer-ho sintàcticament.

Encara que l'experiment mental que proposa Searle no és el més recent, està de plena actualitat, perquè deixa al descobert algunes de les debilitats del funcionalisme, que són fonamentals per la qüestió sobre la IA que ens ocupa. L'experiment queda exposat de la següent manera: imaginem que estem en una habitació i que no sabem xinès. L'habitació està dissenyada de forma que en cada extrem té una ranura. Per una d'aquestes ranures ens van donant caràcters xinesos, als quals hem de respondre per la ranura que es troba a l'altra costat de l'habitació. Certament, no sabem xinès però al centre de la sala hi tenim una taula amb un manual d'instruccions que conté respostes apropiades a qualsevol successió de caràcters xinesos que se'ns faci arribar. Buscant-la al manual, podem donar resposta a cada un dels enunciats que ens donen i comunicar-nos amb la persona que hi ha al darrere de la segona ranura. El manual està ben fet i, per tant, les respostes que dona són correctes. La qüestió aquí és: podem dir que estem parlant xinès? El fet que estiguem responent a caràcters d'acord a un manual que ens proporciona les respostes, no significa ni que entenguem ni que estiguem parlant xinès, tan sols que ens comportem com si en sabéssim. Els enunciats que retornaríem no serien símbols, sinó objectes amb els quals construiríem estructures purament logicoformals gràcies a unes regles molt definides. Tindríem un comportament que exhibiria la capacitat de parlar el xinès, una

competència que, al capdavall, no tindria comprensió. Un fet que ens porta a concloure que l'ús competent del llenguatge no exigeix comprensió.

Amb aquest experiment, Searle vol demostrar que el comportament de la IA és anàleg al que estaríem fent nosaltres en aquesta habitació, evidenciant que la capacitat d'obeir certes regles no implica intel·ligència, comprensió o res vinculat amb la consciència. Des del punt de vista d'un observador extern, ens comportaríem exactament com si entenguéssim el xinès, sense fer-ho en absolut. Admetre que aplicar el mètode de l'experiment no és suficient per a dir que una persona entén i parla xinès, també ens portarà inexorablement a acceptar que cap computador que executi un programa comprengui el xinès, perquè estarà fent el mateix. Aquesta és una conseqüència del fet que no hi ha forma d'obtenir la semàntica a partir de la sintaxi, perquè operar i construir cadenes sintàcticament correctes i ben formulades, no implica saber-ne el significat.

La IA funciona així, segons Searle, donat que és un sistema logicoformal per a formular i revisar cadenes de signes. La relació que les persones tenim amb el llenguatge no és l'execució d'un algoritme per a la manipulació sintàctica de símbols sense cap mena de contingut semàntic. S'estableix una relació diferent quan les persones emetem un enunciat, ja que quan expressem una creença o un desig ens estem referint a objectes i situacions que es donen en el món. Si la ment humana funcionés com un programa informàtic no podríem expressar res que es referís al món. En canvi, creure que hi ha algú davant de la porta de casa meva és establir una relació amb aquesta, una direcció intencional dels estats mentals cap a la realitat. Si els estats mentals i els enunciats informen d'alguna cosa que està passant en el món, és gràcies a la intencionalitat.

El que passa en l'habitació xinesa és que per la persona que està traduint els caràcters amb l'ajuda del manual, aquests símbols no tenen una dimensió semàntica, i per això no comprèn l'idioma, li manca la intencionalitat. Igual que passa amb els sistemes d'IA i la seva falta de comprensió, quan només apliquen el càlcul de probabilitats basant-se en patrons estadístics a partir de la mineria de grans quantitats de dades: sense les dades que els hi proporcionem, no generarien resultats.

En podem posar dos exemples: els classificadors d'imatges i els generadors de música. En el cas dels softwares de classificació d'imatges que et diuen si una imatge és una taula o no, el programa no veu ni potes, ni les superfícies que pugui tenir una taula, ni res de res. El que «veu» són píxels i bits, i estableix patrons estadístics entre els píxels de les imatges en les quals se li ha dit que hi apareix una taula, de forma que quan veu de nou aquests patrons la seva resposta és

que és una taula, però no sap què són les taules. Les aplicacions de generació de música operen d'una forma similar, ja que poden fer obres d'art copiant l'estil d'altres artistes fins a semblar que formen part de la seva obra. Poden accedir a una base de dades amb totes les peces de Bach, per exemple, i fer una simfonia identificant i plasmant les característiques que el compositor barroc donava a les seves obres simfòniques. Però no entén què és la música ni pot emocionar-se en escoltar-la. Tampoc pot crear un estil propi, només copia.

Seria legítim no veure tantes diferències entre aquest funcionament i el que fem les persones com a sistema conscient? El matemàtic Marcus du Sautoy estudia si es podria contemplar que el sistema compost per l'habitació, el manual i la persona són l'entitat que comprèn el xinès (Du Sautoy, 2020, p. 335). Al final, diu, una neurona individual no serà mai capaç de comprendre res, en canvi, el desenvolupament del nostre cervell (neurona a neurona), pot arribar un moment en que sigui capaç de comprendre el llenguatge. D'aquesta forma, quan una persona opera amb els caràcters que rep i el manual, per generar una resposta, en realitat s'està comportant com un conjunt de neurones —que formen part d'un cervell— encarregades d'aquesta funció. Per això, aquesta persona no entén les respostes que està donant en xinès per l'altra ranura, perquè és el conjunt de tot el sistema el que comprèn les respostes que dona. En contra d'això, hem de notar que les habitacions no poden parlar xinès. Les persones que parlen xinès són sistemes, però sistemes biològics —és a dir, vius— capaços de comprendre i parlar xinès, mentre que el sistema format per l'habitació, les ranures i el manual no parla xinès perquè no l'entén. El cervell de les persones que parlen xinès no té una espècie de centre de control on s'activi una zona que manipuli enunciats lingüsitics sense comprensió del seu significat. Per més que hi hagi una persona manipulant aquells símbols i creant cadenes de caràcters —sense entendre-les—, el sistema format per l'habitació i la persona no és un ésser viu que pugui comprendre i aprendre.

Les propietats intencionals i semàntiques de la ment representen un problema pràcticament infranquejable pel funcionalisme. I és que la IA —els teòrics, professionals i desenvolupadors que hi treballen— cau moltes vegades en un formalisme purament sintàctic que obvia absolutament els significats dels símbols. Una forma de solucionar-lo consistiria a proposar una teoria funcionalista de la semàntica, que fos capaç d'explicar en termes causals el significat dels signes amb els quals operem (escrits, parlats i pensats) en relació amb altres signes i la conducta, com han suggerit alguns autors (Block, 1993; Van Gulick, 2009). Tanmateix, convé destacar que suposar que les propietats semàntiques puguin emergir de relacions intrasimbòliques, involucra un emergentisme que presenta el significat com una relació entre significants —sigui amb objectes o amb altres signes—, capaç de reduir tot el contingut semàntic de la realitat a una

amalgama de relacions entre símbols; perquè si el significat de «quadre» és un quadre, els quadres també són símbols. En tal cas, ens seria molt complicat poder reconèixer la realitat, en tant que seria indistingible dels símbols.

En conclusió, el funcionalisme agafa tàcticament una concepció de l'ésser humà impregnada de *naturalisme* quan afirma que les persones són completament definibles des d'un punt de vista científic, perquè d'aquesta manera pot argumentar que el que fan les persones també és reproduïble, incloent la consciència. Amb tot, no proporciona una descripció del que realment és la consciència, tan sols un model, sense arribar a abordar el problema de la consciència en si mateix. Per a no caure en el reduccionisme ment-cervell, és tan fonamental la separació que estableix entre hardware i software, que condueix a un dualisme radical i incoherent perquè resulta problemàtic conciliar la idea d'una realitat material amb la del món de les funcions. La descripció funcionalista de la ment pot donar fàcilment per conscient entitats que realment no ho són, i això dificulta una correcta aproximació del què és la consciència. Que la ment operi d'acord a regles lògiques (algorítmiques), no significa que sempre les hagi de seguir; en canvi, un programa informàtic no pot fer res fora d'aquestes regles. És important, no obstant això, tenir present que la millor identificació que puguem fer de la consciència sempre serà incompleta, ja que no podem traduir tot el volum de dades que conté la realitat a informació i coneixement.

4.7 Conclusions sobre el problema ment-cos

El desenvolupament de sistemes basats en IA ha reobert les consideracions al voltant del problema ment-cos, plantejant si les màquines podrien pensar o fer alguna cosa semblant. En aquest context, explicar com estan relacionades la ment i la matèria és un problema ontològic, i explorar les diverses respostes que s'hi han donat ajuda a comprendre millor que és la consciència i alguns dels prejudicis associats a la IA. S'han presentat les teories sobre les quals es fonamenten els principals pressupòsits en la concepció de la ment i el seu intent de reproduir-la de la neurociència i dels professionals del sector tecnològic.

Hem dit que la neurociència proporciona informació crucial sobre el cervell, però no pot reduir la complexitat de la ment ni l'experiència humana a funcions purament neuronals. És imprecís identificar el cervell amb la consciència perquè, encara que sigui necessari per a la ment, no és pot explicar què o qui és una persona únicament examinant-li el cervell. Davant la possibilitat de que un sistema d'IA sigui conscient, hem argumentat que reproduir el cervell pot significar fer un model de la consciència —en el millor dels casos—, un model que mai serà consciència. La pregunta al voltant de si una màquina pot pensar o desenvolupar consciència s'hauria de

formular de la següent forma: pot desenvolupar consciència un model estadístic? Considerem que fer un model del cervell no implica fer un model exacte de la ment, ja que la comprensió detallada dels processos cerebrals no garanteix una comprensió completa de la ment i la consciència. Creure que un model de la consciència és conscient, és confondre les coses reals amb els models (simplificacions) que podem fer d'elles, és a dir, confondre el mapa amb el territori.

El funcionalisme proposa que la consciència i altres estats mentals poden ser realitzables per diferents sistemes materials, incloent entitats no biològiques com els ordinadors. Per analitzar aquesta qüestió ens hem basat en els criteris desenvolupats per determinar si un sistema d'IA pot ser conscient (Butlin et al., 2023), i hem argumentat que aquests criteris no són concloents. Encara que puguin suggerir que una IA és més probable que sigui conscient com més criteris compleixi, no poden confirmar-ho amb certesa. Al centre d'aquesta problemàtica hi trobem la confusió entre realitzabilitat múltiple i independència del substrat, que pot dur a errors en la identificació de sistemes conscients, especialment en el context de la IA.

La realitzabilitat múltiple significa que una mateixa funció pot ser realitzada per diverses configuracions materials adequades, i això no significa que qualsevol configuració material pugui esdevenir conscient. En canvi, la independència del substrat fa referència a que una funció pot ser potencialment acomplida per objectes que tenen bases materials totalament diferents. De manera que si el funcionalisme considera que un ordinador podria ser conscient, és perquè està aplicant el concepte d'independència del substrat en comptes del de realitzabilitat múltiple.

El funcionalisme no pot explicar els estats mentals fenomenològics perquè hi pot haver sistemes amb disposicions funcionals idèntiques però amb experiències subjectives diferents o absents. L'experiment mental de l'habitació xinesa de Searle mostra que els sistemes que operen únicament segons regles sintàctiques, no poden comprendre realment el significat. Aquesta crítica destaca la limitació del funcionalisme en oblidar la dimensió semàntica i intencional de la ment, dimensió que és fonamental per a la consciència.

CAPÍTOL 5. EL MITE DE LA IA

La IA emergeix com un dels temes que més fascinació desperta del nostre temps, situant-nos en un punt d'inflexió del desenvolupament tecnològic. Encara que aquesta nova era digital sigui certa, la ciència-ficció i el moviment transhumanista alimenten un mite sobre les capacitats de les tecnologies basades en IA, que confon la capacitat per a imitar i simular la conducta humana amb capacitats que són humanes (com la intel·ligència, la consciència o el pensament). El 2023 el Future of Life Institute va presentar una carta oberta en la qual es demanava aturar el desenvolupament de projectes innovadors en IA, per la possibilitat de generar una superintel·ligència (Future of Life Institute, 2023).

Al voltant d'aquesta visió mítica de la IA hi ha un seguit de males interpretacions filosòfiques que convé desmuntar. Els riscos derivats de la IA existeixen, però no tenen a veure amb una tecnologia que ens sotmeti, tenen a veure amb l'ús de caixes negres que desconeixem com funcionen (fins i tot els que les dissenyen), amb la invasió de la privacitat, la gestió de la responsabilitat, etc. Floridi alerta de tres dogmes vinculats a aquesta tendència de sobredimensionar la IA: l'imminent arribada de la superintel·ligència, una tecnologia que sotmetrà a l'ésser humà, i la necessitat d'actuar abans no arribi el tecnoapocalipsis (Floridi, 2022). Amb la intenció de contribuir a aquest desmitificació, aquí ens volem centrar en una lectura crítica del Test de Turing, per aclarir en què consisteix veritablement, i en oferir arguments sobre la impossibilitat que la IA sigui intel·ligent.

5.1 Revisant el Test de Turing: confusions i error

Al plantejar el test, Alan Turing (1950) intenta resoldre el problema de com determinar si una màquina pensa i actua de forma intel·ligent. La seva intuïció és que per aclarir si una computadora té vida mental hem de posar el focus d'atenció en el seu comportament, perquè és l'única via d'accés que tenim per a conèixer els seus estats interns. En conseqüència, desenvolupa un mètode sistemàtic per a trobar proves sobre si el comportament d'una computadora es pot distingir del d'una persona que descriu com a *Joc de la imitació* (*The imitation game*). Si disposem un interrogador humà de tal forma que a través d'un xat hagi de decidir si està conversant amb una persona o una màquina i no és capaç d'identificar a la màquina, aleshores —diu Turing— aquesta ha passat el test. Per descomptat, el paper d'interrogador només el pot fer una persona, un fet que ens adverteix que per Turing l'ésser humà és l'única font de validació dels comportaments humans. Dit d'altra forma, que la tasca de determinar si estem conversant amb un ordinador o amb una persona no és algorítmica, ja que

és una tasca que requereix d'intel·lecte (humà) per a ser resolta. Paradoxalment, en l'actualitat som els internautes els que constantment hem de demostrar que som persones davant d'una màquina que insistentment ens fa prémer el botó «no soc un robot» per validar que som humans.

No hem de restar importància als antecedents que relacionaven el funcionament de les màquines amb el pensament i la intel·ligència, perquè mostren l'obstinada tendència d'atribuir vida mental a aquests dispositius. El matemàtic britànic Charles Babbage —al segle XIX— va dissenyar la que es considera una de les primeres computadores, encara que només la feia servir per fer càlculs. Babbage parlava del comportament de la seva màquina analítica amb expressions com «ensenyar a la màquina a preveure» o afirmant que la màquina «sap». Conscient que estava antropomorfitzant el funcionament del seu artefacte, es justificava dient que només feia servir aquests termes en un sentit figuratiu, per economitzar el llenguatge i estalviar-se llargues expressions (Swade, 2000, p. 103-104), però desconeixia el potencial de la seva invenció. Va ser Ada Lovelace qui es va adonar que la màquina de Babbage —a banda de fer llargs i complexes càlculs— podia tenir molts més usos gràcies a les targetes amb què funcionava, motiu pel qual aquell artefacte es va convertir en la primera màquina programable i Ada Lovelace, en la primera programadora de la història. No obstant això, Ada Lovelace va ser prudent en qualificar el potencial de la màquina de Babbage, veient-hi nombroses possibilitats, però totes limitades a fer només el que se li demanés i sense capacitats creatives. Aquest punt de vista ha derivat en el *Test de Lovelace*, que només és superat si el programa pot fer quelcom per al qual no ha estat dissenyat (Bringsjord et al., 2000).

Inspirades en els seus precursors, les contribucions de Turing el col·loquen com un dels ideòlegs teòrics del que l'any 1956 es va definir com a Intel·ligència Artificial a la conferència *Dartmouth Research Project on Artificial Intelligence*, organitzada per l'informàtic John McCarthy, si bé fins aleshores es feia servir el terme *Intel·ligència de Màquina* quan es referia a al que avui coneixem com a IA. La força del mètode de Turing no està tant en aclarir si una màquina pot pensar, sinó en si una persona és capaç de distingir si una màquina pensa; és a dir, si sembla que pensa. Del que ens informa el Test de Turing és de que si les coses semblants és comporten de forma semblant, els objectes que es comportin de forma intel·ligent també hauran de ser intel·ligents, per tant diu que la intel·ligència i el pensament són una conducta que es pot imitar. Les crítiques que acumula són nombroses, la gran majoria de les quals plantegen objeccions a la definició d'intel·ligència que Turing fa servir i a les condicions que estableix per a

la intel·ligència. Però, aquestes crítiques moltes sovint porten associada dues confusions, comunament acceptades, que poden desviar l'atenció del veritable problema del test.

La primera confusió és que Turing amb el test ofereix una definició d'intel·ligència (ser intel·ligent és comportar-se intel·ligentment), quan realment rebutja entrar a identificar els termes *pensament* i *màquina*[64], perquè considera que discutir sobre aquests significats dificultarà l'anàlisi de la qüestió. La segona confusió és que el test proporciona condicions necessàries i suficients per a la intel·ligència (determinar intel·ligència), quan només planteja —en el millor dels casos— una forma de recollir proves per a decidir si una màquina programada és intel·ligent o pensa (indicis de si una màquina es comporta intel·ligentment). Confusions com aquestes entelen i fan passar desapercebut l'error del criteri de Turing, que és no distingir les qüestions epistemològiques de les ontològiques: considerar que és el mateix que un observador cregui que està justificat a atorgar vida mental a una màquina i que la màquina realment tingui vida mental. A continuació volem justificar per què Turing no aporta una definició d'intel·ligència, contra el que consideren altres autors, perquè no determina condicions necessàries i suficients per a la intel·ligència. Conclourem que el seu mètode alimenta l'estès equívoc que la simulació de conducta intel·ligent és, en si mateixa, intel·ligència (ser no és semblar).

El Test de Turing no pretén definir què és la intel·ligència o el pensament, una intenció que trobem expressada en l'inici de l'article (Turing, 1950) on presenta el test. Rebutja aportar una anàlisi de l'ús comú dels termes *màquina* i *pensament*, perquè comportaria fer una recerca estadística, a l'estil d'una enquesta, per a respondre a la qüestió sobre si les màquines poden pensar. Per a evitar respondre a aquesta pregunta Turing proposa el *Joc de la imitació* i verificar la intel·ligència basant-se en la imitació eficaç d'un comportament.[65] Hauríem de plantejar si el joc assumeix una definició conductual d'intel·ligència, de forma que no la determini a través de conceptes mentalistes, sinó de la competència lingüística exhibida. No podem negar que Turing estableix un correlat entre la conducta lingüística —almenys sintàctica— i la intel·ligència, quedant aquesta última determinada a través de l'anàlisi de la primera: ser intel·ligent és parlar (comportar-se) de forma intel·ligent. Però això no significa necessàriament que Turing pensés el seu test com una forma d'emmarcar o definir la intel·ligència, atès que el lloc central de l'assumpció se situa en les proves obtingudes (empíricament). En una entrevista radiofònica de

[64] Recordem que la pregunta que motiva l'article és al voltant de si les màquines poden pensar «*Can machines think?*» i que llavors es feia servir la locució *Intel·ligència de Màquina* per referir al que ara coneixem com a IA.

[65] Un criteri de validació que inspira el concepte de simulació en IA, d'importància cabdal en la disciplina (González & Vergauwen, 2005).

l'any 1952, Turing deixa veure aquest propòsit quan diu el següent: «es podria dir que aquest test serveix per veure si les màquines pensen, però seria molt millor no formular-lo així i caure en la petició de principi [donar per fet que una màquina podrà pensar si passa el test]» (Copeland, 2001). Ell mateix reconeix que el test només mesura si una màquina és suficientment sofisticada com per simular convincentment que és humana davant les preguntes que se li formulen, cosa que no vol dir que les màquines que superin el test hagin de ser intel·ligents o pensants. Amb això, es demostra que el Test de Turing no estableix una definició sobre què és la intel·ligència, tan sols ofereix un seguit de criteris per avaluar la capacitat d'una computadora per imitar el comportament humà.

Una confusió que es desprèn de l'anterior és que el test determina les condicions necessàries i suficients de la intel·ligència. Amb el que va vist només podem entendre el Test de Turing com una *mena de distinció conductual* (no definició), on l'exigència és que una màquina es comportarà de forma anàloga a la humana si i només si un observador no pot distingir el seu comportament a través d'un examen conductual. Sobre això, Ned Block (1990, p. 249) va més enllà i assegura que aquesta és una definició conductista d'intel·ligència, perquè si bé Turing nega que passar el test sigui una condició necessària per a la intel·ligència, sí que és una condició suficient. D'acord amb la suficiència, això apuntaria que qualsevol entitat capaç de manifestar un comportament indistingible del d'un humà, hauria de ser considerada com a intel·ligent; però també significaria que no tenir tal comportament seria totalment irrellevant per a determinar si és intel·ligent (d'acord amb la no necessitat). De manera que passar el test seria garantia d'intel·ligència, mentre que no fer-ho no revelaria cap informació sobre les capacitats cognitives. Tanmateix, com s'ha anunciat més amunt, Turing només considera que passar el test dona proves (estadístiques o inductives) per justificar que la màquina pugui ser intel·ligent.

Arribats aquí entra en escena la figura de l'interrogador. En el *Joc de la imitació* el què és rellevant és que l'interrogador estigui convençut que està conversant amb una persona i no amb un ordinador. El convenciment de l'interrogador al voltant de que un ordinador pugui ser intel·ligent involucra obtenir proves que així ho justifiquin, i no condicions necessàries i suficients. Turing subratlla la importància que l'interrogador desenvolupa en el joc, posant de relleu que aquest no podria ser expert en informàtica, en psicologia o en ciències de la cognició, ja que fàcilment podria plantejar qüestions intencionadament ambigües o amb dobles significats per a desemmascarar a la màquina.[66] Ens podem plantejar també si aquest interrogador podria

[66] Els experts podrien enxampar a la màquina fàcilment amb tecnicismes o bromes i, en el seu lloc, Turing considera que la persona interrogadora hauria de ser una persona comú, sense més coneixements tècnics

ser un altre computador, dissenyat per a distingir quan parla amb una persona de quan ho fa amb un congènere. Hem avançat la resposta anteriorment, fent menció a què només entitats dotades d'intel·lecte poden validar l'intel·lecte de les altres. A més a més, això ens revela que si no existeix un algoritme que pugui reemplaçar el criteri de l'interrogador, és perquè la intel·ligència humana és irreductible des d'un punt de vista computacional, precisament perquè l'ésser humà sembla ser una peça insubstituïble en el joc. Un programa pot simular ser una persona, convencent un interrogador de la seva intel·ligència; però no pot suplir a l'examinador i el seu criteri per a avaluar què és intel·ligent i què no. El test no aporta condicions necessàries ni suficients per a la intel·ligència, en tot cas són proves per a la validació de comportaments intel·ligents. En canvi, el test sí que ens informa del fet que el rol que desenvolupa l'interrogador no pot quedar recollit en un algoritme.

Havent desarmat les que considerem que són les dues principals confusions que amaguen algunes de les interpretacions que es fan del Test de Turing, examinem l'error que comet Turing amb el seu joc. El test obliga a partir de la premissa que no sabem si estem conversant amb una persona o amb una computadora, però que conductualment ho podem esbrinar. Considera que una simulació *creïble* serà aquella que recolzi la creença de que una màquina és intel·ligent i que té vida mental. Per tant, si la màquina supera el test, el mètode ens porta a creure que qui està conversant és una persona, una creença que és falsa. El que determinarà que un programa conversacional passi el test o no, serà la seva capacitat per produir una creença falsa en els interrogadors, i no que realment tingui vida mental. Podem estar més o menys disposats a acceptar que la simulació de conducta lingüística competent d'un programa sigui adequada per a convèncer als interrogadors de que estan xatejant amb una persona. Però, com es pot adduir que això últim és una prova de que el programa en qüestió pensa i és intel·ligent? La simulació de conducta lingüística no és una prova d'intel·ligència de la màquina, en tot cas seria la prova d'intel·ligència del seu programador. Ser convençut per una computadora tampoc permet justificar que aquesta sigui intel·ligent. L'error de Turing aquí està en confondre l'ontologia amb l'epistemologia, i la tasca amb l'agència.

El problema sobre les altres ments ha estat àmpliament discutit des de la filosofia majoritàriament a través de perspectives ontològiques i epistemològiques, però aquí Turing està barrejant dos aspectes fonamentals sobre aquest assumpte. Una cosa és saber d'altres *ments* a

que la mitjana de la població. Aquesta especificació del perfil de l'interrogador pot semblar prudent i justificada, però en si mateixa comporta haver de trobar un «ciutadà mitjà» i evitar a qualsevol individu amb prou coneixements per destapar a l'impostor.

través de la generació causal de la creença que una màquina és intel·ligent, originada per la resposta conductual similar a la d'una persona. Aquesta és una qüestió epistemològica. En canvi, una cosa diferent és l'existència d'altres ments i de propietats —com la intel·ligència i el pensament—, l'existència de les quals no sembla que pugui dependre de l'anàlisi que un observador extern pugui fer o de les proves que puguin justificar a un interrogador a creure que un ordinador té vida mental.[67] Aquestes darreres són qüestions ontològiques i tenen a veure amb que una màquina sigui realment intel·ligent. De la generació causal (a través de l'observació, per exemple) d'una creença, no se segueix que l'objecte de la creença sigui cert. De l'observació d'un comportament que associem amb la intel·ligència no se segueix l'existència d'intel·ligència. El que pretenia Turing era reduir la pregunta sobre si les màquines pensen mitjançant l'elaboració d'un mètode de verificació que corroborés l'associació de conductes a estats mentals, però en fer-ho Turing confon l'epistemologia amb l'ontologia. D'aquesta forma, Block té raó sobre que el Test de Turing és un mètode conductista, però s'equivoca respecte que és una definició conductista d'intel·ligència i que aporta condicions suficients sobre la intel·ligència.

Una cosa és el criteri segons el qual podem reconèixer la intel·ligència i una altra és la intel·ligència en si mateixa. El filòsof italià Maurizio Ferraris (2022) crítica aquest error, conegut com la *fal·làcia transcendental*,[68] consistent a derivar l'ontologia de l'epistemologia, tot i que admet que és una confusió molt natural. El que ens diu és que l'origen d'aquesta fal·làcia està en confondre el que hi ha i allò que sabem o que creiem saber. Si ens aturem a considerar les implicacions metafísiques d'aquesta fal·làcia, —assegura Ferraris— veurem que suposa una pressuposició superficial sobre l'existència de ments independents de la matèria amb la capacitat de produir representacions i, fins i tot, l'objecte d'aquestes. En última instància, això ens aporta arguments per a defensar que les simulacions de la realitat no són la realitat, de la mateixa manera que imitar el comportament d'una persona no et fa ser aquella persona.

Alguns autors (Dreyfus, 1994; Floridi, 2015; Searle, 1980) subratllen que, encara que el Test de Turing associï conductes amb pensament, el problema filosòfic no es limita exclusivament al fet que la conducta lingüística és insuficient per a validar vida mental, sinó en que la simulació o imitació (entesa com a font de prova estadística i element fonamental del test) tampoc permet concloure categòricament que una computadora pensi o sigui intel·ligent. El problema del joc de la imitació es troba en que atribuir pensament a un ordinador programat per a simular

[67] Una postura que és defensada per Daniel Dennett (1989) amb el seu funcionalisme.

[68] Una confusió molt comuna, similar a l'*error d'estímul* segons el qual és fals que quan tanquem els ulls no estem veient res, donat que realment estem veient coses tals com fosfens i altres imatges.

convincentment estats mentals és confondre *imitar* amb *ser*. Imitar un comportament no implica que qui imita i qui és imitat comparteixin estats mentals (ni condició ontològica), en la millor situació s'estaran duplicant els comportaments, però no els estats mentals de l'individu imitat. Convé destacar que els enfocaments que s'acullen a la identitat entre ser i imitar, consideren que una imitació és intel·ligent si convenç a algú de que ho és. De forma que l'èxit de la simulació dependrà de la persona que hagi de validar si aquella imitació s'ajusta als seus criteris de ser intel·ligent, la petició de principi de la que parla Turing.

El plantejament de Turing no només s'emmarca dins el conductisme, també s'adscriu en el funcionalisme, que considera que la vida mental pot ser reduïda al funcionament d'un programa que es realitzi en una multiplicitat de disposicions materials. És contra aquest supòsit que Searle formula l'experiment mental de *L'habitació xinesa*, per a demostrar que la manipulació de símbols sota regles preestablertes no es tradueix amb que parli xinès, tot i que des del punt de vista d'un observador extern es podria concloure —basant-se en el comportament— que és un competent parlant de xinès. El mètode de Turing, segons Searle, només és un forma conductista de verificació de vida mental, que alimenta els punts de vista dualistes i que, sense cap justificació, assumeix que la intel·ligència i el pensament es poden replicar a través de la simulació (Searle, 1980, 1985). Com hem vist, la simulació convincent d'intel·ligència, que només pot passar a un nivell epistèmic, no necessàriament ha d'implicar intel·ligència. Sota el model de Turing, se separen radicalment els factors responsables de la intel·ligència —que són biològics— de la mateixa intel·ligència.

5.2 La IA no és intel·ligència artificial

Al voltant de la reflexió sobre si la IA reuniria funcionalitats que estan associades a les capacitats humanes (consciència, intel·ligència, pensament), molts dels arguments que es presenten en contra se situen en una escala epistemològica: intentar demostrar que l'ésser humà posseeix unes capacitats que li manquen a la màquina. S'estableix així una competició entre home i màquina (subjecte-objecte) per veure qui és més llest, qui resol millor els problemes, qui es desenvolupa amb major solvència en entorns canviants, qui és capaç de sentir, qui comprèn el que diu, etc. En el marc d'aquesta discussió i sense treure rellevància als arguments que despullen a la IA d'aquestes capacitats, convé assenyalar que la intel·ligència, el pensament o la consciència tindran sentit sempre que estiguin encarnades en un sistema biològic, que ha de respondre a unes necessitats adaptatives que tenen a veure amb la vida i amb saber-se situar en el món. A continuació, el que es vol demostrar és que els sistemes basats en IA, realment, no són

intel·ligència artificial, si entenem *intel·ligència* com quelcom similar a la intel·ligència humana i si entenem *artificial* com a adjectiu d'aquesta intel·ligència.

Es pot parlar d'intel·ligència separada de l'organisme o hardware que l'encarna? Aquesta és una pregunta que ens situa a la frontera entre el que els sistemes d'IA poden i no poden fer, perquè si se suposa que la intel·ligència és la capacitat per a resoldre problemes —com creuen els professionals del camp de la IA— hi podem respondre d'una forma; en canvi si creiem que el sistema ha de complir altres característiques —a banda d'aquesta— hi respondrem d'una altra manera. Són múltiples les dificultats que trobem per a definir amb precisió la noció d'intel·ligència. Prova d'això és que, tot i haver estat estudiada amb profunditat per diverses disciplines (psicologia, sociologia, política o biologia), difícilment podem formular una definició d'intel·ligència que l'esgoti de forma transdisciplinària. No hi ha un únic criteri per a la intel·ligència, sinó múltiples, la qual cosa ens alerta que no estem davant d'un concepte rígid i propi de la ciència: la intel·ligència sempre és múltiple.[69]

La idea d'intel·ligència amb què treballen els professionals del sector tecnològic és estrictament logicoformal perquè queda abstreta dels materials que la conformen.[70] Vista així la intel·ligència es redueix a la capacitat de fer operacions lògiques —desmaterialitzades i descorporitzades—; a seguir una successió de regles per a trobar la solució a un problema. Es pot entendre aquest algoritme per a solucionar problemes com la capacitat de fer sil·logismes o de treure conclusions. Els relats que conceben la ment com un programa informàtic i els programes informàtics com a ments, contribueixen a desencarnar la intel·ligència i la confonen amb un formalisme més. De fet, la IA està encarnada i materialitzada perquè depèn de recursos naturals, consumeix energia, requereix mà d'obra i necessita ocupar un espai físic. Per tant, aquest substrat material existeix perquè la tecnologia computacional necessita manipular estats de dispositius electrònics, encara que moltes vegades es passi per alt aquest detall. Kate Crawford ha assenyalat que aquesta visió de la intel·ligència és completament desencarnada

[69] La *Teoria de les Intel·ligències Múltiples*, proposada per Howard Gardner (2019, p. 24), representa una visió pluralista de la ment, que reconeix moltes facetes diferents de la cognició, té en compte que les persones tenen diferents potencials cognitius, contrasta diversos estils cognitius i defineix set tipus d'intel·ligència (lingüística, logicomatemàtica, espacial, musical, corporal, interpersonal i intrapersonal (més endavant afegeix la naturalista al considerar-la essencial per a la supervivència de l'espècia humana).

[70] L'investigador d'IA de Google, Blaise Agüera, considera que «la immensa majoria de models de llenguatge demostren per primera vegada que el coneixement del llenguatge i la intel·ligència poden desassociar-se de totes les característiques corporals i emocionals que compartim entre nosaltres i amb molts més animals» (Agüera, 2022).

(*disembodied*) de qualsevol relació amb el món material, i que el que convindria és reencarnar la intel·ligència (Crawford, 2021, p. 7).

Realment, la intel·ligència no es pot situar fora d'un sistema biològic —com un software— i és inseparable d'un cos —com una ment aïllada—, perquè és relacional: es desenvolupa i es manifesta a través de la interacció amb l'entorn i amb altres éssers, no hi ha intel·ligència sense relacionar-se en el món (Colbert et al., 2020). En l'enfocament relacional es posa de manifest que el pensament, la resolució de problemes i la presa de decisions són processos que es configuren a través de les relacions amb altres persones, amb la societat i amb el context cultural, així com amb les eines o tecnologies que fem servir. I aquesta capacitat per a relacionar-se amb l'entorn és un aspecte que té tanta importància com la mencionada part més logicoformal de la intel·ligència a la qual recorren —sovint de forma exclusiva— els professionals de la IA. Els agents intel·ligents necessiten un cos per poder tenir experiències vivencials directes del seu entorn, donat que depenen de les seves capacitats sensorials i motores per a poder interactuar-hi. Sense un cos, aquestes representacions de l'entorn serien abstractes i no tindrien contingut semàntic. És gràcies a la interacció directa amb l'entorn que els agents poden relacionar els senyals a través dels seus sensors amb representacions simbòliques generades a partir del que han percebut (López de Mántaras, 2018).

Creure que les màquines poden pensar, aprendre o ser intel·ligents és una actitud que queda reflectida en el vocabulari que es fa servir en el camp de la computació i que, com hem vist, ja es feia servir per descriure el comportament de les computadores més primitives. Quan es parla de *Machine learning*, s'està fent referència a que és la màquina la que està aprenent, caient en una antropomorfització dels sistemes que fan servir aquesta tecnologia. És convenient corregir aquesta presumpció perquè, en tot cas, si algú està aprenent alguna cosa, són les persones gràcies a la utilització d'aquestes eines per a automatitzar tasques ben definides (identificar patrons, fer càlculs, etc.). No hi ha problemes per a l'algoritme que ens assisteix a resoldre'n un perquè tampoc té intencionalitat. En tot cas el dissenyem, li proporcionem les dades, l'entrenem i l'executem per a nosaltres interpretar els resultats que n'obtinguem. En el fons, pel disseny i desenvolupament tecnològic, que l'algoritme sigui intel·ligent o no, és completament irrellevant, el que interessa és la interpretació que es faci de les conclusions que proporcioni. La seva intel·ligència queda relegada al disseny i l'ús que se'n faci.[71] I quan ens referim al seu ús no només pensem en la manera (com) i en la finalitat (per a què) amb què es faci servir o les conseqüències

[71] Un fet de gran rellevància per abordar les qüestions ètiques, i que situa a la persona com a agent responsable i no la tecnologia, com veurem en parlar de l'agència ètica.

que es puguin derivar dels seus usos (responsabilitat), sinó també al fet que sigui utilitzat (la seva implementació activa dins d'un sistema o procés).

Que un sistema d'IA estigui ben dissenyat i ofereixi conclusions valuoses, no tindria impacte real si no s'aplica en un context determinat. La seva intel·ligència, o valor potencial, només té sentit si se'n fa ús, perquè és a través del seu ús que genera efectes reals, siguin positius o negatius. En el marc de la IA, únicament té sentit parla d'intel·ligència en tant que el sistema s'insereixi i interactuï amb persones; d'altra forma, si no es fa servir per algú i no afecta a ningú, ni tan sols es pot dir que el sistema estigui funcionant. De la mateixa manera que un llibre que no és llegit per ningú no compleix la funció d'un llibre (fer gaudir, comunicar coneixement o experiència), només en virtut de que un sistema d'IA es faci servir per —o afecti— a algú podrem parlar d'intel·ligència. El sentit de les tecnologies basades en IA no recau només en la seva capacitat aïllada per a analitzar o processar dades, sinó en la seva aplicació pràctica i en els efectes que té sobre les persones i els entorns en què s'insereix. Si té algun sentit parlar de la intel·ligència d'un sistema d'IA, aquesta no es troba en les seves capacitats per a establir patrons estadístics, sinó en com aquestes capacitats són interpretades, utilitzades i afecten a les persones. És en l'ús actiu, en la interacció amb el món, que la IA *adquireix* sentit i rellevància. Així, si haguéssim de parlar de la intel·ligència en el marc de la IA, aquesta és relacional, depenent de la seva relació amb els humans, els contexts i els resultats que genera.[72]

Amb freqüència la intel·ligència s'associa a una sort evolutiva que ha privilegiat a sistemes biològics complexos per a respondre de la forma més adequada possible a les situacions que se'ls presentin. Altres éssers vius no humans són també intel·ligents, però existeix una diferència entre la intel·ligència d'aquests éssers i la dels humans. Els éssers humans segurament som els únics éssers vius de la terra que saben que són éssers conscients de la terra i que saben que la terra és un planeta que gira al voltant d'un astre que gira al voltant d'una galàxia. Comunament es dona per vàlid que la intel·ligència humana és natural perquè ha estat fruit d'aquest procés transformador i adaptatiu. En contraposició, la intel·ligència de la IA, que —en l'intent per emular la intel·ligència humana— és artificial, al ser un artefacte que no ha comptat amb una evolució biològica que l'hagi fet més adaptable.

No obstant això, la intel·ligència humana no és completament biològica o natural, si bé no existiria si els humans no fóssim éssers vius. La intel·ligència humana també és un artifici de

[72] Una idea que porta associada una reflexió ètica: com més depenem de sistemes d'IA en la presa de decisions, més rellevant esdevé com es fan servir, qui els fa servir i amb quina finalitat.

l'ésser humà, perquè depèn completament del context cultural, social i polític. En l'intent per explicar què i com és l'ésser humà —i renegar també d'aquesta autodescripció—, s'inventa una nova forma d'intel·ligència que busca més enllà d'un mateix i que aconsegueix transcendir la finitud de la seva naturalesa l'animal humà. Es tracta d'una intel·ligència que es perpetua a través d'institucions educatives que s'encarreguen de transmetre els coneixements a les noves generacions perquè els qüestionin, els integrin i els ampliïn. En la mesura en què s'estén —en l'espai i en el temps— el coneixement amb la transmissió cultural i el sistema educatiu diem que la intel·ligència humana és artificial (supera allò biològic i individual), perquè es converteix en un instrument del nostre entorn cultural i del qual no en pot escapar.

La intel·ligència no es pot alliberar del seu context per ser concebuda de forma aïllada d'una història, una cultura, un entorn social o, fins i tot, d'un subjecte. Però, tampoc pot quedar reduïda als cervells. Ja assenyalaven Clark i Chalmers (1998) que els processos cognitius no es troben tots al cervell. Una intel·ligència mancada d'un cos, presentada com una substància immaterial, és un contrasentit perquè no existeix res tal com les intel·ligències separades. Agafant l'exemple del Test de Turing, es comet un error quan es fa servir aquest test com a criteri d'intel·ligència i es conclou que si és superat, hem d'estar davant d'un ésser intel·ligent al nivell humà. Amb ocurrència, el pensador Gustavo Bueno (Bueno, 1996, p. 140) —rellevant per les seves aportacions al materialisme filosòfic— manifesta que el que s'hauria de postular en el Test de Turing és que la computadora desenvolupés un cos humà tan bon punt satisfés el criteri, ja que aquesta és l'única forma a través de la qual podria donar forma a un *logos*, una raó o una consciència.

En darrer terme, el Test de Turing falla perquè una màquina digital és logicoformal i incapaç d'entendre la semàntica, de dirigir-se cap a alguna cosa que no sigui ella mateixa (*inputs-output*), motiu pel qual podem dir que és incapaç de comprendre. En les persones els continguts intencionals dels estats mentals no depèn en exclusiu de propietats autoreferencials, també depenen de la seva comunitat lingüística i social. Per tant, el que fixa el contingut intencional dels estats mentals, també té a veure amb les relacions que establim amb altres individus conscients. D'aquí la importància de la col·laboració social per tal de determinar en què pensen les persones. El control sobre els nostres pensaments, les nostres creences i desitjos no és fruit d'una visió intel·lectualment aïllada del seu contingut, sinó que en part depèn del benefici social que aporta compartir llenguatge i l'atribució d'estats mentals a altres individus. En el Test de Turing es parteix de l'habilitat sintàctica de la màquina, per a demostrar que la computadora té

estats mentals, un supòsit que demana el principi que vol demostrar perquè pressuposa que una màquina sense un cos pot emular un ésser humà.

No es tracta únicament que l'intel·lecte humà no es pugui capturar en termes logicoformals, sinó que és externalitzat mitjançant artefactes, com les institucions educatives, els llibres, les escoles i la cultura, que transcendeixen les capacitats subjectives de l'individu.[73] Les persones dipositem part del nostre intel·lecte en artefactes que construïm, però és només quan entra en joc la nostra interacció amb aquests que podem parlar d'intel·ligència. Encara que la seva existència sigui continuada i independent de nosaltres, no significa que les capacitats que tenim gràcies a un llibre, una calculadora o un mòbil persisteixin en ells, fins i tot, després del seu ús. Les reflexions de Bruno Latour són especialment rellevants quan suggereix que amb l'avenç tecnològic, el límit que separa les intel·ligències humanes de les artificials s'ha diluït: «amb la introducció de tantes tecnologies intel·lectuals […], s'ha desdibuixat la pròpia distinció entre les intel·ligències naturals, situades i tàcites, i les artificials, transferibles i desencarnades» (Latour, 1996, p. 300-301). Els artefactes tecnològics encapsulen capacitats intel·lectuals, però són desencarnats, en el sentit que només adquireixen significat quan són utilitzats pels humans.

La tecnologia, tal com assenyala Latour, no és intel·ligent per si mateixa, sinó que la seva intel·ligència està situada i activada en el context de la interacció humana. La intel·ligència humana es demostra en la fabricació de ginys, que depenen de l'intel·lecte humà per a poder operar. Un mòbil intel·ligent es presenta com una entitat intel·ligent per si mateixa, com si el seu intel·lecte s'exhibís de forma incondicionada i autònoma. A ningú li sembla estranya la idea de que un mòbil intel·ligent —que incorpora el calendari, el correu electrònic i les xarxes socials— és una extensió de la ment. El mòbil permet recordar i planificar el dia a dia, contestar missatges i estar informats del que passa al món, sense haver de pensar en quan hem d'anar a la cita amb el metge, en contestar una carta o en veure als nostres amics per comunicar-nos-hi; en altres paraules, amplia allò que pot fer la nostra ment per si sola. En aquest context, les paraules del filòsof David Chalmers en prologar el llibre *Supersizing the Mind* d'Andy Clark (2011) cobren més sentit que mai: «quan algunes parts de l'entorn són acoblades amb el cervell de la manera apropiada, arriben a formar part de la ment».

[73] Es podria argumentar que la IA també és social i política (perquè treballa en xarxa), o que en l'experiment de l'habitació xinesa de Searle el conjunt del sistema (habitació, manual, persona) és intel·ligent, però en cap dels casos podem dir que la intel·ligència s'externalitzi perquè mai transcendeix ni se situa fora del conjunt de d'elements que conformen la IA o l'habitació.

Hem de tenir en compte que, per una banda, la part intel·ligent del sistema no es troba en el sistema, donat que depèn de la manipulació humana per a ser-ho (no és intel·ligent per si sol, sinó una extensió del giny humà); i per l'altra, la part que conté el sistema no és intel·ligent (no hi ha intel·ligència sense un context), perquè l'algoritme ha estat dissenyat per seguir unes regles sense comprendre-les i li manca el cos per tenir vivències (biografia). Si entenem la intel·ligència com alguna capacitat similar a la humana, i artificial com el que emana de la intervenció humana, els sistemes d'IA no són intel·ligents i només són artificials com a creació humana. En cap cas s'hauria d'entendre aquesta artificialitat com a adjectiu d'una suposada intel·ligència computacional o mecànica, perquè aquest atribut fa referència al fet que aquests artefactes, sistemes i components tecnològics són fets per persones. En conclusió, és pertinent precisar que la IA no és intel·ligència artificial.

5.3 Conclusions sobre el mite de la IA

El mite de la IA consisteix en sobreestimar les capacitats de la tecnologia basada en IA i està a l'arrel de les concepcions distorsionades d'un futur apocalíptic. A través d'aquest mite hem analitzat de quina manera una mala interpretació del Test de Turing i la suposició de que la IA és intel·ligent contribueixen a tenir una visió esbiaixada de les capacitats i possibilitats de la tecnologia.

La interpretació del Test de Turing sovint confon la simulació de comportament intel·ligent amb l'existència real d'intel·ligència. Superar el test implica només que una màquina pot imitar la conducta humana de manera convincent, no que aquesta màquina posseeixi capacitats mentals o intel·ligents. El Test de Turing no proporciona una definició precisa d'intel·ligència ni estableix condicions necessàries i suficients per a aquesta. En comptes d'això, serveix com a mètode per avaluar si una màquina pot simular comportaments humans, deixant de costat la seva ontologia i la realitat dels seus estats mentals. Amb la formulació del test, Turing confon aspectes epistemològics i ontològics en la seva anàlisi, assumint que la creença d'un observador sobre la intel·ligència d'una màquina pot justificar que sigui realment intel·ligent.

Existeix el mite que la IA és intel·ligent, una capacitat que tan sols podem atribuir a éssers biològics (amb la necessitat d'adaptació al món per sobreviure). La intel·ligència es presenta com un concepte múltiple i relacional que va més enllà de les simples operacions logicoformals que sovint s'associen a la IA. Mentre que la intel·ligència humana és inseparable del context biològic i social (biogràfic) en què es desenvolupa, la IA és un producte creat per humans que no posseeix les mateixes capacitats. Els sistemes d'IA poden realitzar operacions complexes i analitzar dades,

però això no implica que tinguin una comprensió real del món que els envolta, ja que els seus processos són purament mecànics i no intencionals.

A més, en tant que la intel·ligència humana no és només biològica, sinó que també és un producte cultural que es transmet a través d'institucions educatives i artefactes, l'intel·lecte humà transcendeix allò biològic i individual. Només podem parlar d'intel·ligència en el marc de la IA quan és aplicada i interactua amb humans, però per si sol un sistema d'IA no pot ser considerat intel·ligent.

Els riscos derivats de la IA no es troben tant en la creació de màquines superintel·ligents com en la manera en què utilitzem aquestes tecnologies i la nostra comprensió limitada sobre el seu funcionament. Els sistemes d'intel·ligència artificial sovint es perceben com a agents capaços de prendre decisions de forma autònoma, la qual cosa ha fomentat la creença errònia que aquests sistemes poden arribar a ser conscients o actuar amb intencionalitat. Els autèntics perills i riscos relacionats amb la IA tenen a veure amb l'ús de sistemes que funcionen com caixes negres en la presa de decisions que afecten la vida de les persones, sense una comprensió adequada del seu funcionament i sense mecanismes adequats per garantir la responsabilitat i la transparència.

CAPÍTOL 6. ÈTICA EN LA IA

Pensar sobre la IA des de l'ètica significa que des de la filosofia ens hem de plantejar qüestions fonamentals sobre l'ontologia (què és pròpiament la IA?) i l'epistemologia (com podem verificar si la IA sap o coneix res?) de la IA. Contestar a aquestes preguntes és al que ens hem dedicat durant els capítols anteriors, a fi de construir una imatge del que és la IA. En aquest capítol ens volem dedicar als aspectes ètics de la IA, tenint present les reflexions ontològiques i epistemològiques anteriors, que ens ajudaran a orientar la discussió al voltant de què és l'ètica de la IA, en quin sentit l'ètica es pot integrar en una tecnologia, si podem parlar d'agència ètica d'una màquina i quins usos tecnològics hauríem d'evitar.

6.1 A què ens referim quan parlem d'ètica de la IA?

L'ètica de la IA ens situa en el si de la reflexió al voltant de l'ètica tecnològica, una preocupació que inicialment van posar de manifest Mario Bunge en definir el terme *tecnoètica* (Bunge, 1977) i Hans Jonas amb l'*ètica aplicada a la tecnologia* (Jonas, 1979), i que també podem vincular amb el context que va motivar el sorgiment de la *bioètica* a mans de Van Rensselaer Potter (1970) com a disciplina que reunís en un sol àmbit el coneixement científic (els fets) i l'humanístic (els valors). La qüestió està en que, si bé, la bioètica ha gaudit d'una sistematització (principis i mètodes) per a la resolució de problemes, la tecnoètica o ètica tecnològica, com anota (Diéguez, 2017, p. 62), encara es troba en una etapa preliminar del seu desenvolupament. Mentre que els darrers anys han proliferat neologismes que posen el cognom *ètica* a les diverses aplicacions tecnològiques (*roboètica*, *ètica digital*, *nanoètica*, *infoètica*, *ètica algorítmica*, etc.) amb la intenció de posar sobre la taula la reflexió sobre les seves implicacions (morals, ètiques i polítiques), això no ha estat necessàriament acompanyat per un augment real de l'orientació per a resoldre els problemes pràctics que planteja la tecnologia (Sætra & Danaher, 2022). Trobem gairebé tantes formulacions d'una ètica tecnològica com subcamps i altres noves tecnologies, fet que indica que l'ètica de la tecnologia està posant més atenció en la mateixa tecnologia que no en la finalitat que persegueix (telos). D'aquí el repte i interès per contribuir en aquest terreny.

La tecnologia no és neutra. Es diu que la tecnologia *a priori* no és ni bona ni dolenta, que depèn dels usos que se'n faci. Gràcies a Ortega y Gasset i Heidegger sabem que aquest supòsit és fals, perquè si bé és una eina, no és neutra. Certament, només les persones poden actuar bé i malament, i en conseqüència podem fer servir les eines per a fer el bé o per a fer el mal. El fet rellevant és que la tecnologia es troba immersa en una xarxa en la qual hi intervenen moltes mans (*many hands*) (van de Poel et al., 2012) i no queda limitada únicament a un conjunt d'eines,

sinó a un context social, cultural i polític. I aquest context no és axiològicament neutre. Els artefactes tecnològics porten incrustats de *sèries* uns valors i unes intencions (no vol dir que tinguin intencions, només que el seu disseny i implementació persegueix unes intencions). En el cas de la IA, i en especial de les caixes negres, el seu grau de sofisticació posa de manifest que aquests valors i intencions poden arribar a ser opacs fins i tot pel mateix desenvolupador del programa.

És necessària una ètica tecnològica perquè la tecnologia ni és neutre ni és imparcial, sempre serveix a una moral —més o menys— concreta. Al capdavall, la tecnologia és un mitja lligat a una teleologia fixada per les persones (desenvolupadors, empreses i usuaris), per tant, abans que res es tracta d'una eina que hauria d'estar al nostre servei si volem que sigui ètica. La qüestió, assenyala Torralba (2022, p. 15), es troba més en com garantir que la tecnologia estigui al servei del benestar social i no només al de les empreses. Les tecnologies no s'insereixen únicament en les nostres vides com un bé o un servei, són *moral materialitzada* (P. P. Verbeek, 2008) perquè porten implícites les actituds i els valors de les persones que les desenvolupen —un fet que als mateixos professionals pot passar inadvertit—, en tant que ens diuen com les hem de fer servir i fan una distinció entre correcte/incorrecte o bé/mal. Motiu pel qual és un requisit fonamental reclamar que el comportament de les tecnologies estigui alineada amb les intencions i valors humans, ja que no fer-ho comportaria riscs significatius (Ji et al., 2024).

L'ètica de la tecnologia es presenta com una disciplina pragmàtica que comprèn problemàtiques derivades de les tecnologies i les seves aplicacions. Però, donat que el domini d'aquesta disciplina ha de poder integrar conceptes tecnocientífics i filosòfics, es deriva una dificultat que té a veure amb trobar un llenguatge comú sense simplificar una a l'altra. Una solució a aquesta dificultat proposaria traslladar la fórmula dels Comitès d'Ètica i Bioètica als Comitès de Tecnoètica, aplegant a persones amb diverses formacions acadèmiques i professionals per a resoldre els problemes ètics derivats dels usos tecnològics en les seves activitats professionals (Echeverría, 2010). Luciano Floridi també recorre a la bioètica quan ofereix una sortida de l'esmenada dificultat a través dels quatre principis bioètics (beneficència, no-maleficència, autonomia i justícia), als quals incorpora el principi d'explicabilitat en el cas de l'ètica de la IA (Floridi et al., 2018; Floridi & Cowls, 2019), una idea que ha estat criticada qüestionant la suposada necessitat d'un principi addicional (Cortese et al., 2022; Loi et al., 2020).

En una de les seves facetes, tractar els aspectes ètics de la IA és preguntar-se per l'impacte (si bé la reflexió ètica va més enllà dels impactes) que aquesta tecnologia té en la vida de les persones. La integració de sistemes basats en IA en el nostre dia a dia ha possibilitat que cada

cop més artefactes puguin prendre decisions de forma autònoma (menor grau d'orientació humana) sense la necessitat que hi intervinguem persones. S'ha transformat la manera en què ens relacionem i també en què ens autopercebem. Per això, l'ètica de la IA no s'ha de centrar exclusivament en les persones (usuàries i desenvolupadors), sinó que també ha d'abordar qüestions que afecten a la relació entre les persones, qüestions que tenen a veure amb la gestió i l'organització de la presa de decisions. Saber què pot fer la tecnologia ajuda a fonamentar les qüestions ètiques que es deriven d'aquestes funcionalitats. Tanmateix, la pregunta ètica no consisteix a saber si els sistemes d'IA poden conduir cotxes, prendre decisions o què és allò que poden fer, sinó que se centra en quines haurien de ser les raons per a delegar tasques a la IA, si ho haurien de fer i de quina forma.

Si l'objectiu que persegueix la IA és imitar el comportament humà —en un intent per reproduir la intel·ligència, la consciència o la ment—, això de bon inici planteja problemes ètics propis de l'ésser humà relacionats amb els biaixos cognitius, l'engany, la confiabilitat i la responsabilitat. Si la IA s'ha d'assemblar a com actuen les persones, seria imaginable —si més no com a possibilitat— que també compartís els seus prejudicis (M. Boden et al., 2017). L'ètica de la IA es fixa en les diverses facetes i actors involucrats en el desenvolupament tecnològic, per estudiar-ne els usos, els impactes i responsabilitats al respecte. És en aquesta interacció que la IA també esdevé una eina d'interès per a la política, donat que s'han de coordinar les esferes socials, econòmiques i normatives en les quals es troba immersa, i també perquè la seva aplicació convida a repensar els problemes tradicionals de la filosofia política (la llibertat, la igualtat, la justícia, la democràcia i el poder) (Coeckelbergh, 2023). En aquest sentit, la urgència de definir una ètica per a orientar la resolució de problemes al voltant de la IA és també una qüestió de prudència, de prevenció de riscos i maximització de beneficis. En definitiva, es tracta d'una qüestió política, perquè només en el si de la política es pot donar l'ètica.

Entesa com a disciplina filosòfica, l'ètica es defineix com la reflexió crítico-racional al voltant de les morals (Román, 2016) i, per això, la denominació *ètica de la IA* ens sembla excessiva perquè no hi pot haver una ètica única i específica per a la IA. Avaluem l'ètica en funció de les seves finalitats, i les finalitats les avaluem des de les activitats humanes on s'insereix (des del context). Com s'està començant a assenyalar, l'ètica al voltant de la tecnologia sembla haver-se desorientat, abordant anècdotes morals més que problemes pròpiament ètics (Pareto & Torras, 2024).

6.2 Qüestions ètiques des de l'ontologia i l'epistemologia

L'ètica de la IA també s'ha de preguntar per quina és la contribució de la IA en el marc de les qüestions ètiques, fet que ha generat un destacable debat sobre la capacitat d'aquests sistemes per intervenir en la resolució de problemes que tradicionalment les persones han abordat a través de la reflexió ètica. Volem contribuir a aquest debat a la llum dels aspectes ontològics i epistemològics sobre la IA abordats, destinant aquesta secció a analitzar quatre conceptes que tenen rellevància en el marc de l'ètica: problema, percepció, coneixement i dubte.

Pensem perquè tenim problemes. Podem entendre que un problema té a veure amb quelcom que impedeix desenvolupar la voluntat i que, més enllà d'un problema cognitiu, és un obstacle per a exercir l'autonomia. En tant que entenem la tecnologia com una eina, donat que es relaciona amb l'intent de resoldre o assistir-nos en la resolució de problemes, si pretenem que la IA ens ajudi en la presa de decisions dels nostres problemes (perquè recordem que la IA no té problemes), haurem de decidir en quins problemes ens poden ajudar sistemes basats en IA, en quin tipus de problemes no podrà o no hem de deixar que ens ajudi i en quins serà desproporcionat recórrer a aquesta tecnologia per a resoldre'ls (l'ús de la IA té un impacte ecològic que s'haurà de ponderar per a determinar aplicacions innecessàries).

La percepció és rellevant per a la qüestió al voltant de quins problemes pot ajudar-nos a resoldre una eina equipada amb IA. Començant pel fet que una màquina no pot percebre, perquè la percepció està relacionada amb la vivència i en veure's a un mateix situant-se en el món (com hem vist en parlar de la percepció des d'Aristòtil i Merleau-Ponty a l'apartat 3.5). Les persones ens obrim a un món que està constituït per més coses que dades, no només captem la realitat sinó que a més a més la sentim i ens involucrem en ella. Això és, ens vinculem amb el món de forma transcendent. Tanmateix, hem d'acceptar que el que fa la IA és un model de la percepció: sense arribar a percebre o sentir-la, capta la realitat d'alguna manera. D'aquestes limitacions epistemològiques es deriva que la IA no pot atendre els contextos (no són computables) en els quals estan imbricades les persones, perquè, malgrat ser bona establint patrons estadístics, la IA no ens pot ajudar en la recerca del sentit de la vida o la llibertat.

El nostre coneixement és limitat perquè l'ésser humà és un ésser limitat. Quan prenem decisions i proposem solucions als nostres problemes sempre hi ha aspectes que s'escapen del nostre control i que no podem dominar (variables indisponible). Quan contemplem la realitat ho fem des d'una perspectiva, des d'un punt de vista. Hem d'abandonar la concepció hegeliana d'una causa completa a través de la qual ho podrem tot perquè haurem aconseguit explicar i

conèixer-ho tot, com també ambicionava la ciència moderna i que ara queda representada per un positivisme dataista. És en la impossibilitat de controlar-ho tot on apareix la vida, que s'escapa constantment del nostre domini i ens sorprèn, emociona o atemoreix. Si conèixer és acumular i processar dades (Big data), hem resolt tots els problemes o ens hem quedat sense problemes.

Però en la IA ja no hi ha perspectiva des d'on es mira el món, només dades. Tampoc hi ha una obertura sintent a la realitat que ens pugui sorprendre, ni tampoc hi ha vida. En aquest context, deleguem les nostres decisions a sistemes d'aprenentatge automàtic que poden analitzar grans quantitats de dades en pocs instants i donar una resposta, pensant que aquesta anàlisi proporciona una resposta basada en coneixement. Però el cert és que la resposta es basa en un model estadístic que res té a veure amb el coneixement necessari per a prendre decisions relacionades amb la vida (com a context biogràfic). Es pot utilitzar la tecnologia per a desenvolupar nous coneixements, això no implica que la tecnologia conegui, sinó que pot ser emprada com a mètode perquè les persones coneguin. Tota tecnologia s'ha de mantenir com una eina o un mitjà a través del qual resoldre els nostres problemes, mai com a un fi per a resoldre'ls (eliminar-los) tots.

Dubtar és fonamental per articular una resposta ètica. Ens permet qüestionar les nostres pròpies creences i considerar que podríem estar equivocats. Un algoritme no té dubtes, perquè el correlat computacional de dubtar seria inoperable, simplement resol problemes per a les persones. Això ens porta a una reflexió: la presa de decisions és una activitat lligada a la llibertat humana i, per tant, no és computable.

6.3 Pot ser ètica la IA?: límits i responsabilitat

Al respecte de si les màquines poden integrar o estar equipades amb un codi ètic, l'enginyera informàtica del MIT Rosalind Picard (1997, p. 19) ha apuntat que com més lliure (autònoma) sigui una màquina, més necessari seran unes normes morals per a controlar-la. Una afirmació com aquesta ens recorda que la relació entre persones i sistemes d'IA ha d'estar mediada per la supervisió i responsabilitat humana, a fi de que les màquines treballin amb valors alineats amb els dels humans. Però el motiu no es troba en que una màquina hagi de comportar-se èticament, sinó en el fet que les persones dissenyen els artefactes i interactuen (entre persones) a través d'ells. En l'assaig *Màquines morals* (*Moral Machines*), Wallash i Allen (2009) estableixen els fonaments sobre com desenvolupar una espècie de moral computacional. Un plantejament com aquest posa de manifest la tendència de desplaçar a les persones del focus d'atenció dels temes que aborda l'ètica de la IA, per centrar-hi la mateixa tecnologia. Una tendència sobre la qual

descansa la premissa que, si molts sistemes d'IA poden operar de forma més ràpida i eficient que les persones, perquè no hauríem de considerar-les responsables del que fan també (Lin et al., 2014, p. 87; Müller, 2020).

En aquesta línia, hi ha iniciatives que persegueixen l'objectiu de dotar als sistemes d'IA de principis per a orientar el seu procés de presa de decisions, de forma que aquest estigui basat en alguna cosa semblant a la reflexió ètica (Gordon, 2020). Per dur a terme aquesta empresa s'estan utilitzant variacions de les estratègies de processament de la informació de dalt a baix (processament guiat per conceptes) i de baix a dalt (processament guiat per les dades). En l'estratègia de dalt a baix es vol desenvolupar un programa específic que permeti a una màquina l'abordatge de problemes ètics. En aquest cas, s'entrena a una xarxa neuronal amb un seguit de conflictes ètics amb les seves respostes —que assumirem temporalment que poden tenir una resposta comprensible pel sistema—, sota el supòsit que, a través d'aquesta base de dades, el sistema podrà respondre a nous conflictes ètics que se li presentin (Guarini, 2006). En la variació de l'estratègia de baix a dalt, s'ha optat perquè sigui a través de les teories ètiques deontològiques i utilitaristes que l'algoritme prengui les decisions (Dehghani et al., 2011). La idea està fonamentada en estudis empírics (de les troballes en psicologia sobre la presa de decisions) per crear un model (MoralDM) del que les persones decideixen en casos concrets i, així, identificar-ne els valors. D'aquesta manera, el sistema aplicarà respostes utilitaristes, sempre i quan els valors no entrin en conflicte, i donarà respostes deontològiques quan existeixi algun conflicte de valors.

En el primer plantejament (de dalt a baix) el problema està en que les possibilitats de situacions reals sempre superaran qualsevol llistat de conflictes amb la seva resposta, de forma que se'ns revela que l'ètica no és una col·lecció de dades amb un mètode per a respondre a un *input*: una prova de que no es pot reduir l'ètica a un algoritme. Cal remarcar que l'ètica no és un procés mecànic o matemàtic que pugui ser modelat amb unes regles fixes aplicables a qualsevol situació, sinó que involucra interpretació, judici i comprensió contextual. L'ètica no només tracta les qüestions sobre la presa de decisions, també reflexiona sobre els valors subjacents, la història, la cultura i les relacions humanes, aspectes que varien segons les circumstàncies (context). Els problemes ètics no tenen una única resposta, sinó que estan envoltats per una complexitat contextual que només els humans poden percebre i atendre. Un altre punt clau és que els algorismes operen amb patrons estadístics i regles predefinides, mentre que l'ètica exigeix flexibilitat i creativitat.

En el segon enfocament (de baix a dalt), a més del que hem criticat al primer, és conflictiu pretendre que el sistema d'IA modelitzi què és prendre una bona decisió en un context determinat, basant-se en la recerca empíric. L'empirisme és conservador perquè es basa en el passat i dificulta transcendir els fets coneguts (el futur es comportarà de forma similar al passat), de manera que a través de la recerca científica en psicologia, en la millor de les situacions, es modelitzarà l'opinió de la majoria al voltant de què és prendre una bona decisió. Conèixer les decisions que l'ésser humà pren en donades situacions, no diu gaire sobre com és el procés de presa de decisions i no diu res en absolut sobre què és actuar èticament. És molt discutible que aquest sistema modelitzi com a correctes aquelles decisions alineades amb l'opinió majoritària. No és a això al que es fa referència quan es parla d'IA alineada amb valors, sinó a l'exigència que els sistemes d'IA han d'adherir-se a les normes morals de la comunitat per estar socialment orientats (Ji et al., 2024, p. 49). Respondre sempre favorablement als interessos de la majoria en molts casos no serà equitatiu —portant conseqüències indesitjables— i, per sobre de tot, serà èticament reprovable. També és criticable la possibilitat de copsar tot el significat de les ètiques deontològiques i teleològiques en termes de distribucions estadístiques (que és el que poden fer els sistemes d'IA), com si un algoritme les pogués entendre.

En qualsevol cas, el debat al voltant de l'agència dels artefactes tecnològics compta amb les veus d'enginyers, científics i filòsofs en la identificació de quin estatus tenen entitats com els robots (socials, assistencials o industrials), els sistemes d'IA, els vehicles autònoms, etc. Existeixen estudis que parlen d'*agència moral* (Floridi & Sanders, 2004) o altres que fan servir indistintament *agència ètica* i moral (Müller, 2020), sense atendre a la distinció entre ètica i moral.[74] Conseqüència d'això és que es confereixi agència ètica a entitats de forma arbitrària, i que el que fan poc té a veure amb l'agència ètica. El nostre plantejament és que només les persones poden ser agents ètics. Per això, examinarem quin estatus tenen les eines tecnològiques i per quin motiu no cauen sota la categorització d'agent ètic.

6.4 Agència ètica i algoritmes

L'estatus d'agent ètic s'acostumava a atribuir a subjectes racionals, apel·lant a al fet que, en tant que éssers racionals, coneixen les decisions que prenen, perquè són accions autònomes (lliures), i, per tant, els subjectes racionals són responsables de les decisions que prenen i de les seves conseqüències. Si la intel·ligència dels sistemes d'IA té a veure amb aquesta racionalitat,

[74] Floridi (2013) també introdueix el concepte d'*acció moral distribuïda* per explicar que hi ha accions moralment neutres de diversos agents que poden resultar en una acció moralment significativa.

els hauríem de dotar de *certa* agència ètica. Així ho creu el l'expert en IA i vicepresident de Google Research, Blaise Agüera (Agüera, 2022), que aplicant un enfocament conductista a l'estil del test de Turing, argumenta que xarxes neuronals artificials complexes com GPT-4 (el model de llenguatge que utilitza ChatGPT) podrien ser considerats persones. L'extreballador de Google, Blake Lemoine, va fer públic que el model de llenguatge que estava desenvolupant (LaMDA) era conscient o que, com a mínim, es comportava com si ho fos[75] (Degli-Esposti, 2023, p. 16).

Hem ofert arguments en contra de que aquest funcionament que suposadament demostren tecnologies com ChatGPT o similars, tingui res a veure amb la consciència, la intel·ligència o la comprensió. Els models de llenguatge operen amb distribucions estadístiques de dades que els permeten interactuar exitosament a nivell sintàctic, però no a un nivell semàntic. No comprenen absolutament res del significat dels textos que generen o de les preguntes que se'ls hi fan, només associen paraules i locucions en funció de la freqüència amb què apareixen vinculades entre elles. També hem vist que els sistemes d'IA tenen moltes dificultats per treballar amb escassetat de dades, incertesa o contradiccions i que just allà on les persones responen prenen decisions, un algoritme respon amb inoperància. A més a més, el problema de la detenció (examinat a l'apartat 3.3) ens anuncia que no tots els algoritmes poden ser computats, perquè hi ha aspectes de la vida humana com el dubte o l'ètica que no es poden reduir a establir una relació entre patrons.

S'ha fet palesa la tendència humana a antropomorfitzar el seu entorn, una inclinació que també pot contribuir en l'explicació de per què conferim l'agència ètica a entitats inertes que no són ni intel·ligents ni sintents. Descartes ja va contribuir en destapar aquest biaix al plantejar que els animals són màquines mancats de la capacitat de sentir, perquè la racionalitat està en sintonia amb la sensibilitat (capacitat de sentir) i negar-ne una implica negar l'altra. Descartes va establir una separació ontològica entre les persones i la resta d'animals, però avui sabem que molts altres animals no humans són sintents. Que els animals són màquines és una idea que s'ha abandonat i, contràriament a Descartes, autors com Singer (1987) consideren que, en tant que el animals senten i la sintonia que existeix entre sensibilitat i racionalitat, els animals mereixen consideració ètica (analitzar l'abast de les nostres accions sobre els animals perquè són criatures que poden patir). Sota aquesta perspectiva, l'estatus d'agent ètic queda determinat per la capacitat de patir que té un ésser. Altres autors discrepen que els animals siguin agents ètics, ja que —si fos el

[75] Que es comportava de la forma en que es comporten les persones quan diem que entenen el que fan i els motius pels quals ho fan: quan són conscients.

cas— haurien de tenir la capacitat de poder i decisió per què se'ls tractés com a tals (Bueno, 2006).

La premissa que els animals són agents ètics perquè senten ha despertat veus que insisteixen en que els algoritmes, robots i programes també poden sentir i que, per tant, mereixen la condició ètica (McGinn, 1993). Al capdavall, en el passat els animals havien estat despossetts de sensibilitat i va resultar no ser així; i ara passa el mateix amb els artefactes tecnològics, potser només falta temps per descobrir que realment senten. Desafortunadament pels que defensen això, el temps encara no ha conferit sensibilitat als robots[76], un escenari diferent és que amb el desenvolupament tecnològic futur arribin màquines que puguin sentir (un fet que aquí ni neguem ni afirmem). La diferència fonamental està en què compartim moltes característiques amb la resta d'animals no humans, característiques que ens venen donades com a éssers biològics que som i que no compartim amb els sistemes d'IA. Els animals estan fets de matèria viva, de carn i són fruit de l'evolució biològica, com nosaltres. En canvi, els sistemes d'IA estan fets de matèria morta, de silici, i són fruit del disseny d'alguna persona.

Assenyala Coeckelbergh (2021, p. 56) que, no obstant això, maltractar un robot és moralment reprovable. Defensa que és així perquè atacar a un robot degrada la nostra integritat ètica, no perquè el robot senti i se l'estigui danyant, i que en virtut d'això se l'hagi de tractar de tal forma que s'eviti el seu patiment (com es fa amb els animals). Per contra, l'antropòloga i experta en ètica de robots i IA, Kathleen Richardson, ha promogut els drets de les màquines juntament amb la prohibició dels robots sexuals, perquè considera que amb la integració de sistemes d'IA és una forma de perpetuar la prostitució i l'esclavatge sexual (Richardson, 2016). Seguint aquesta estela, es podria defensar la prohibició de videojocs violents perquè estem atemptant contra la dignitat dels personatges no jugadors (NPC per les sigles en anglès *non playable character*) que hi apareixien. Independentment de la conclusió a la qual s'arribi, l'argument no s'hauria de centrar en si aquestes màquines o entorns virtuals tenen capacitat de sentir o no, sinó en l'impacte que tenen en les relacions humanes. Ens ha de preocupar que aquestes tecnologies perpetuïn estereotips i comportaments que afectin a les persones reals, no que es degradi la integritat moral d'un robot o d'un entorn virtual. Igual que els videojocs violents no es prohibeixen perquè els NPC pateixin, no hauríem de prohibir els robots sexuals pel fet que puguin experimentar alguna mena d'emoció. En tot cas, hauria de ser perquè contenen una visió excessivament simplificada de les relacions humanes i dels rols de gènere, fomentant dinàmiques de dominació

[76] Entesa com la capacitat de sentir, sí que la tenen si l'entenem com la capacitat de comportar-se diferentment davant un estímul donat, com hem vist.

i cosificació. Així doncs, la regulació d'aquestes tecnologies hauria de posar el focus en els efectes que tenen sobre la societat i les vides de les persones, no en la tecnologia en si mateixa.

Reprenent l'experiment mental de l'habitació xinesa que planteja Searle, de la conducta no en podem derivar la consciència. Que algú respongui correctament a les preguntes que se li fan en xinès, o que expressi que sent dolor o felicitat —perquè té un manual de la resposta precisa que ha de donar—, no es suficient per a inferir que aquest entengui el significat del que se li pregunta i que contesti o senti el dolor o la felicitat que diu sentir. Els humans podríem mentir al respecte, cosa que no deixaria de significar que comprenem el que diem o que sentim el que expressem (que som conscients), encara que el motiu pel qual s'arribés a concloure això fos diferent. Però, un sistema d'aprenentatge automàtic, només *reconeix* el que volen dir les paraules o l'expressió de les emocions com a producte d'haver estat entretant amb una infinitat de textos etiquetats com a exemple d'un significat o emoció i la freqüència amb què apareguin relacionades en els textos que se li presentin.[77] En el cas de robots socials com en Pepper, que poden reconèixer emocions, no significa que coneguin el que és el dolor o la felicitat, sinó que han estat entrenats amb tantes imatges que poden establir patrons estadístics entre les imatges que se'ls han mostrat i expressions de dolor i de felicitat.[78]

Més enllà de la capacitat que tingui un sistema per sentir, la consideració d'agent ètics per alguns autors (Tomasik, 2014) es pot fixar per la capacitat d'aprendre per reforç, i no tant per la capacitat per a sentir. Sota aquesta concepció serien agents ètics les persones, els animals i també diversos sistemes d'IA i, fins i tot, s'ha fundat una associació per promoure tractar èticament als algoritmes que aprenen per reforç (*People for the ethical treatment of reinforcement learners*[79]). L'argument és que, en última instància, l'ésser humà i els animals són sistemes d'aprenentatge per reforç (assaig i error), essent aquest el mecanisme que explica la nostra forma de privilegiar conductes que ens són beneficioses i rebutjar aquelles que ens són perjudicials. El senyal de recompensa per a un algoritme d'aprenentatge per reforç és anàleg al plaer i al dolor per als sistemes biològics.

[77] Sigui dit que etiquetar no és universalitzar, perquè mai recollirà la individualitat

[78] Un problema associat a aquesta identificació estadística de les emocions a través del reconeixement facial és que, per exemple, normalment quan les persones somriuen és que estan experimentant felicitat, mentre que hi ha estudis que vinculen el somriure en persones que pateixen un trastorn d'espectre autista amb experimentar estrès o ansietat (Rump et al., 2009). És una dificultat que s'ha de tenir en compte en robòtica social i que ens alerta de que la generalització estadística desatén el cas particular.

[79] Veure més al web http://petrl.org.

D'aquesta forma la bondat de les accions queda definida per criteris logicoformals, donat que les persones no són altra cosa que un algoritme d'aprenentatge per reforç que ha estat implementat en la carn i no en un xip. En el marc d'aquest enfocament, en un programa informàtic aquest algoritme d'aprenentatge per reforç està materialitzat pel hardware, en canvi en el cas de les persones ens hauríem de preguntar a on està inscrit aquest algoritme. No és que es pugui reduir la forma que té l'ésser humà d'aprendre a un codi que quedi recollit en el cervell o en els gens; perquè la forma d'aprendre no és un algoritme computable, donat que depèn tant del cervell, com dels gens, com del cos, com de la voluntat, com de la cultura i les eines que fem servir per a compartir-la.

Floridi i Sanders (2004) consideren agents ètics als sistemes d'IA —encara que no tinguin llibertat, consciència, responsabilitat o capacitat per sentir— pel fet que interactuen amb les persones, són autònoms i s'adapten a l'entorn. És en aquesta capacitat per a captar estímuls, no necessitar supervisió constant i poder ajustar les regles per a complir un objectiu, que poden desencadenar accions danyines o beneficioses —fer bé o mal—, i per això consideren que tenen agència ètica. El qüestionable del seu enfocament està en què assumeixen una autonomia del sistema en un sentit massa ampli, ja que l'autonomia d'un sistema d'IA es pot vincular en funció de la independència del control humà, però no hi ha autonomia fora de les capacitats per a les quals se l'hagi programat: no podrà fixar objectius propis, perquè li venen donats, ni podrà modificar -los. És per aquest motiu que en un sentit ètic, l'autonomia només es pot atribuir a les persones —i no a als programes o robots— perquè són les que programen, dissenyen, fan servir i controlen aquests algoritmes i, per tant, les que en són responsables.[80] En conseqüència, si un sistema d'IA que causa mal per errors o funcionaments defectuosos, qui haurà de respondre d'aquests danys seran les persones que l'han desenvolupat, programat i implementat. Altrament, no tindria sentit esperar responsabilitzar a un algoritme per contenir biaixos racistes o condemnar a un robot per fer allò pel qual estava programat fer o per un mal funcionament.

La raó per la qual no podem acceptar que un algoritme sigui un agent ètic és, per un cantó, ontològica i, per l'altra, epistemològica. Un sistema operatiu no és una entitat sintent, perquè del fet que es comporti com que sent —reproduint els comportaments de les persones quan diem estar contentes o tristes— no se segueix que senti —que estigui content o trist—. Els patrons estadístics no senten. Per tal de saber el que és causar dany o plaer és necessari (no

[80] És d'interès en aquest debat tenir en consideració els sis tipus d'eines tècniques que es poden utilitzar per a explicar els diversos nivells d'autonomia tècnica: manual, maquinal, automàtica, autonomia tècnica integrada, semiautonomia tècnica i autonomia (Funk & Coeckelbergh, 2020).

suficient) haver-ne experimentat, així és com es defineix la *paciència ètica* (Véliz, 2021). Si no hi ha comprensió del patiment, manca la paciència ètica, i si no hi ha paciència ètica tampoc hi pot haver responsabilitat.[81] Som responsables perquè experimentem dolor i podem entendre el dany que infligim.

A l'estil de la crítica al test de Turing, podem pensar en un programa que imiti el comportament de les persones que considerem ètic, ara bé, que es comporti èticament no vol dir que ho sigui, no vol dir que reflexioni sobre l'abast de les seves accions (correlació no implica causalitat). Seria com reduir tota la reflexió ètica de l'ésser humà a una conducta, sense atendre els motius pels quals s'orienta la presa de decisió en una direcció o una altra. En ètica no es posa l'atenció tant en el *què*, sinó en el *per a què*. Un programa que ha modelitzat el que és actuar èticament, podrà exhibir una conducta fonamentada en aquest model, ignorant el per a què es comporta d'aquella manera perquè no es capaç de sentir.

Sumàriament, hem vist que fent referència a la capacitat de sentir, l'aprenentatge per reforç i la imitació de conductes ètiques, l'agència ètica no es pot atribuir als robots, algoritmes o programes basats en IA, ni tampoc la paciència. La forma en que ens relacionem amb els artefactes tecnològics no és en el si d'una interacció entre individus, en tot cas entre persones i eines. Sovint es cau en la il·lusió que estem *conversant* amb un model de llenguatge perquè el que fa és similar al que fem els humans quan xategem o que *juguem* contra programes informàtics. Més aviat el que estem fent és interactuar amb una màquina, i no sempre ho fem en igualtat de condicions, perquè per més que diem que Kaspàrov i Sedol van ser vençuts pels programes Deep Blue i AplhaGo respectivament, seria més encertat dir que els que van guanyar van ser els equips d'enginyers i programadors que van dissenyar els sistemes.

També hi ha una discussió al voltant de la confiabilitat de la IA (*trustworthy AI*), mentre que només són mereixedores de confiança les persones perquè estan dotades de sensibilitat i es poden responsabilitzar del que fan (Ryan, 2020). La confiança no es diposita en els sistemes d'IA, talment com no ho fem en un pont, es confia en les persones que han dissenyat i fet els ponts i els sistemes d'IA i en la fiabilitat de la seva feina. Així mateix, de l'única manera que podem parlar d'ètica quan parlem d'IA, és en el context de la relació entre un humà amb un altre humà, perquè els sistemes d'IA no poden ser ètics al no ser agents ètics. Qui s'ha de responsabilitzar si aquestes

[81] La paciència és condició necessària per a l'agència, però no suficient. Per això els animals mereixen consideració ètica, perquè saben que poden patir, tot i que no són agents ètics perquè no es poden fer càrrec (responsables) del dany que poden fer, són pacients ètics i prou.

tecnologies no funcionen correctament i provoquen algun dany, són els agents ètics que les han desenvolupat. Mentre que els sistemes tecnològics que ens assisteixen poden prendre algunes decisions que afecten a les persones, la responsabilitat continua essent humana per permetre-ho.

6.5 Contra les caixes negres

Si s'ha de permetre que els sistemes d'IA prenguin decisions que tindran implicacions en les vides de les persones —paciències afectades per la tecnologia—, és fonamental comprendre com funcionen abans de fer-ho. Amb aquesta finalitat, també ens interessa abordar des de l'ètica els problemes vinculats a les caixes negres, les quals funcionen de tal forma que és molt complicat saber perquè, donada una entrada, donen una sortida. La problemàtica respecte les caixes negres i les dificultats per a conèixer com funcionen, adopta dos significats diferents (Wachter et al., 2018). En un sentit, la mateixa complexitat de les profundes xarxes neuronals impedeix que els programadors i experts en IA puguin comprendre per què un sistema està funcionant d'aquella manera. En un altre sentit, es fa referència al fet que els mateixos usuaris no siguin conscients de com funciona la tecnologia ni de l'abast de les seves accions a través dels seus usos.

Els programes inicials d'IA, dissenyats seguint l'estratègia deductiva (de dalt a baix) per a processar les dades, eren transparents perquè pels programadors era senzill conèixer i controlar de quina forma l'algoritme *aprenia* per a realitzar una tasca molt concreta. Els seus processos de presa de decisions es podia entendre i analitzar, permetent veure com les entrades es converteixen en sortides. En són exemples els arbres de decisió, que mostren el camí que segueix l'algoritme per arribar a una conclusió, o les regressions lineals, que presenten les relacions entre les variables de forma clara i comprensible. En ambdós casos és possible saber perquè es donen els resultats que es donen, en funció d'unes regles de decisió i els valors assignats a les dades d'entrada. Amb la implementació d'estratègies inductives (de baix a dalt), són les dades les que van orientant les regles que seguirà aquell algoritme. Els programadors dissenyen els sistemes presentant-los enormes quantitats de dades i exemples, perquè aquests generalitzin un model de la tasca que se li està demanant que faci. Aquests sistemes poden ser empleats per fer multituds de tasques, però els mecanismes que utilitzen queden ocults. Es converteixen en caixes negres perquè no podem saber en què s'han basat per arribar a una conclusió.

És convenient que ens preguntem si estem —o si hauríem d'estar— disposats acceptar conviure amb artefactes que no sabem com funcionen, als quals tanmateix els hi confiem les

nostres vides. Al cap i a la fi, no saber com funciona una tecnologia no ens impedeix introduir-la a la societat: agafem el cotxe sense saber mecànica, volem amb avions sense saber aeronàutica, fem servir el microones sense saber què són les ones electromagnètiques i ens connectem a internet amb els mòbils sense saber com funcionen les ones de ràdio. La diferència està en que els mecànics i enginyers que han construït aquests artefactes sí que entenen els mecanismes que els composen, i són ells els que hauran de respondre si per algun error de funcionament causen algun dany. La transcendència que han suposat els sistemes d'IA és que ni els mateixos encarregats de desenvolupar-los coneixen com treuen conclusions. Estem davant d'un esdeveniment sense precedents, que planteja als professionals del sector tecnològic —i també als dels sectors que utilitzen aquestes eines— la qüestió: com fer-se responsable del procés de presa de decisions que fa un sistemes d'IA que no sabem com funciona?

Aquesta pregunta ens l'hauríem de fer pel que fa els resultats indesitjables d'un sistema i el mal que poden generar, i també ens ho hauríem de plantejar quan les decisions que prengui el sistema siguin adequades i donin una resposta funcional adaptada a la situació. No fer-ho seria com ignorar els biaixos que pot contenir, però, sobretot, seria assumir que la IA ens ha superat i rendir-nos davant la seva màgica capacitat per a treure conclusions. Acceptar els resultats d'una tecnologia sense saber com funciona ni posar-los a prova, és conferir-li una divinitat que no necessita més explicació, és claudicar davant el funcionalisme.[82] No ens podem imaginar a un pilot d'avió confiant-li la seva vida, la de la tripulació i la del passatgers a un sistema d'IA que no sap com funciona, en qualsevol cas hauria de saber perquè fa les coses que fa. El mateix passa amb els algoritmes de conducció autònoma, de finances o de diagnòstics mèdics.

La qüestió ètica és que no es poden deixar en mans de circuits opacs les decisions que afecten la vida de les persones. Per més que els resultats que obtenim d'un sistema que funciona com una caixa negre siguin els òptims, és irresponsable acceptar acríticament les decisions que ofereixi. El programador d'una xarxa neuronal d'aprenentatge profund hauria de poder explicitar quin és el motiu pel qual el sistema ha proporcionat un resultat i no un altre. No conèixer els mecanismes que el fan funcionar, diu Coeckelbergh (2021, p. 103), significaria que en cas de mal

[82] Contràriament a això, l'expert en ètica de l'IA, Scott Robbins pensa que els requisits de transparència només suposen un problema quan els resultats són negatius, perquè si un sistema d'IA fos molt eficient en la detecció d'algun tipus de càncer de forma incomprensible per a nosaltres, el valor d'aquesta informació superaria els problemes derivats de la incertesa de saber com hi ha arribat a fer el diagnòstic (Robbins, 2019).

funcionament la responsabilitat hauria de ser de la persona que hagi decidit fer servir un sistema tan opac en comptes d'un més transparent.

El principal problema que es planteja aquí és que, com hem vist, com més autonomia tècnica té un algoritme, més complexes són les tasques que pot realitzar, però més fàcilment es pot convertir en una caixa negre i perdre el control sobre la forma com modelitza una bona presa de decisions. Els sistemes que tenen més capacitat per a detectar patrons i predir comportaments, són justament els que resulta més complicat desxifrar com funcionen. En canvi, els sistemes que més fàcilment podem explicar perquè han arribat a una conclusió i no a una altra, acostumen a ser també els que tenen menys èxit en l'anàlisi de dades i en les seves prediccions. Amb l'augment de l'autonomia també argumenten les tasques que pot acomplir un sistema, tanmateix, no per això és més responsable. Tal dissociació genera el que Matthias (2004) qualifica com *escletxa de responsabilitat* (*responsability gap*), que posa especial atenció en què passa si no es pot predir el comportament d'una màquina.

Una forma que tenim per atribuir responsabilitat és a través de la condició de control i la condició epistèmica (Fischer i Ravizza, 1998). Així, la responsabilitat es pot fixar en funció del grau de control que es tingui sobre un comportament i del coneixement que es tingui del que s'estigui fent (saber el que es fa i ser conscient d'estar-ho fent). Sobre la condició de control, és responsable qui té el control. En el cas de màquines autònomes que funcionen amb aprenentatge profund, veiem que les persones poden perdre fàcilment el control sobre els comportaments d'aquestes. Donada aquesta possibilitat s'hauria d'exigir que els humans sempre poguessin assumir el control, anul·lant, si fos necessari, les decisions del sistema. Pel que fa decisions que s'han de prendre en lapses de tremps molt curts, existeix la possibilitat que aquesta transferència de control no es pugés dur a terme. Si són decisions que suposarien un dany potencial per la vida de persones, en la línia de Coeckelbergh (2024, p. 101), podem suggerir que aquestes aplicacions mai es desenvolupin. Sobre la condició epistèmica, és responsable qui sap el que està fent. Aquesta condició estableix que els desenvolupadors han de saber com funcionen els seus ginys per ser responsables, fet que requereix transparència i coneixement. I alhora, també diu és rellevant per atribuir responsabilitat conèixer la tecnologia que s'està utilitzant. Tant els desenvolupadors com els usuaris han d'estar informats per ser conscients de l'abast de les seves accions i les limitacions de la tecnologia. Davant les caixes negres és complicat satisfer la condició epistèmica perquè es desconeix de quina forma un sistema arriba a una decisió. A més a més, es dificulta exponencialment poder informar com funciona una tecnologia si tenim en compte que hi ha tants professionals implicats en el seu

procés de fabricació que ells mateixos desconeixen aspectes rellevants del seu disseny i funcionament. Hi ha tantes mans implicades en l'elaboració d'un sistema d'IA, que és problemàtica l'atribució de responsabilitat en el sentit que, tant pels programadors com pels usuaris, sempre hi ha alguna cosa que desconeixen del seu funcionament (van de Poel et al., 2012).

Per últim, creiem que aquestes condicions (epistèmica i de control) vinculades a la responsabilitat són aspectes que haurien de ser limitants a l'hora de desenvolupar noves tecnologies. La responsabilitat, ja ens advertia Hans Jonas (1995), s'articula en el marc de la necessitat que l'ésser humà té per actuar amb prudència davant l'enorme potencial transformador que posseeix la tecnociència. Advertir els perills i els riscos que estem disposats o no a assumir, ens revelen que la rendició de comptes no només és sobre quines persones són responsables i de què són responsables. Hi ha una dimensió que té a veure amb les persones que resultaran afectades (els pacients), per la qual cosa també hem de parlar sobre davant qui som responsables (Smith, 2015). Si s'ha de tenir control sobre les accions i conèixer el que s'està fent, és per poder respondre davant els afectats, els desenvolupadors i les empreses han d'estar disposats i ser capaços d'explicar què i perquè alguna cosa ha fallat (Coeckelbergh, 2024, p. 104). Per tant, la rendició de comptes (respondre davant d'algú) també està molt relacionada amb la condició epistèmica, i això fa que també sigui complicat acatar amb aquesta exigència si es treballa amb sistemes que no es comprèn com funcionen com les caixes negres.

6.6 Conclusions sobre l'ètica en la IA

La reflexió ètica al voltant de la IA parteix de la premissa que la tecnologia, inclosa la IA, no és ni neutra ni imparcial. Els sistemes d'IA estan immersos en un context social, cultural i polític, porten implícits valors humans (moral materialitzada) que influeixen en les seves aplicacions. Sovint la reflexió tecnoètica es desvia massa cap a la tecnologia i cap a les finalitats que persegueix (teleologia). En aquest sentit, l'ètica de la IA ha de centrar-se en com els sistemes tecnològics serveixen o afecten a les persones i la societat de la qual formen part, i no en la tecnologia en si mateixa. La reflexió ètica s'ha de centrar en l'impacte en les persones, no únicament en la tecnologia.

Si parlem d'ètica de la IA no és en virtut de que les computadores es puguin comportar èticament o haguem d'esperar que ho facin, sinó perquè les persones es relacionen amb altres persones a través d'aquesta tecnologia. Parlar d'una ètica exclusivament aplicable a la IA ens sembla innecessari, perquè les preguntes que pot suscitar el desenvolupament de la IA també

involucren qüestions que afecten altres tecnologies. Aquestes qüestions poden ser abordades des de l'ètica com a disciplina única, amb múltiples objectes d'estudi (tecnologia, robòtica, IA, etc.). Això ens permet reflexionar sobre els mateixos principis aplicables a diverses àrees tecnològiques i contextos.

Més enllà de que els sistemes d'IA puguin incloure quelcom semblant a un algoritme per a la reflexió ètica, els enfocaments presentats demostren que no aconsegueixen crear un bon model algorítmic del que l'ésser humà fa quan reflexiona sobre qüestions ètiques. No obstant això, l'ètica no es pot reduir a un algoritme o a una col·lecció de dades. L'ètica implica interpretació, comprensió contextual i reflexió, aspectes que són inherents a la complexitat humana. No és possible codificar un conjunt complet de regles que pugui abordar totes les possibles situacions morals. L'ètica implica judici, comprensió i adaptació a contextos que un sistema d'IA no pot captar ni processar de manera significativa.

Si entenem que un agent ètic ho és perquè pren decisions de forma lliure i que, per això, és responsable de les conseqüències que comporten aquestes decisions, els sistemes d'IA que obeeixen unes regles establertes no tenen agència ètica. Si l'agència ètica té a veure amb la capacitat per sentir, els artefactes tecnològics tampoc en tenen. Si bé tenen sensibilitat a alguns estímuls, la capacitat de patir no es reduïble a certa forma de processar les dades, sinó que té a veure amb estar viu. Encara que un sistema d'IA estigui programat per tal de comportar-se com si experimentés dolor, gràcies a l'experiment de l'habitació xinesa de Searle podem entendre que exhibir aquest comportament no es tradueix en tenir l'experiència del patiment: els patrons estadístics no senten.

Que no puguem atribuir agència ètica als sistemes d'IA implica que hauran de ser les persones les responsables dels mals resultats o mals usos d'aquestes tecnologies, d'altra forma no tindria sentit responsabilitzar a un algoritme d'haver estat mal programat. Els sistemes d'IA no poden assumir una agència ètica pròpia perquè operen amb patrons estadístics i regles predefinides, i no tenen la capacitat d'entendre o percebre el context moral en el qual es desenvolupen les accions.

Hem analitzat si hauríem de permetre que sistemes com les caixes negres —que donen respostes sense que puguem saber en què es basen per fer-ho— prenguin decisions que afectaran a la vida de les persones. S'han presentat les condicions de control i epistèmica per a defensar que s'ha de poder mantenir control sobre les tecnologies que es desenvolupen i que

s'ha de conèixer com funcionen (tan per part dels usuaris com dels desenvolupadors, però sobre tot per part d'aquests darrers, que seran qui informaran als primers).

A la llum d'aquests criteris, els sistemes d'IA que funcionen com a caixes negres haurien d'estar limitats i no aplicar-se quan el seu mal funcionament pugui posar en perill la vida de les persones. El motiu és que seria complicat saber quan això podria passar al no conèixer de quina forma estan prenen les decisions. No poder predir de quina forma es comportarà una xarxa neuronal artificial, significarà que en cas de mal funcionament el responsable hauria de ser qui ha decidit fer servir un sistema tan opac.

Una ètica basada en el funcionalisme i el computacionalisme ignora el context i les complexitats de l'agència humana. Per prendre decisions ètiques és necessita agència i l'agència és entendre el context, cosa que el funcionalisme sovint desatén. A més, enfocaments com aquests redueixen la conducta humana a patrons predictibles, allunyant-se d'una comprensió completa dels valors. No podem dotar a les màquines de característiques epistèmiques o ètiques comparables a les humanes. No obstant això, la IA pot contribuir en el debat ètic, sempre que els humans en mantinguem el control i la responsabilitat, assegurant-nos que la seva aplicació sigui transparent i respectuosa amb els valors humans. En conclusió, la IA pot ser ètica, però en tant que ho siguin les persones que la dissenyen i les finalitats amb que es faci servir, no per si sola la IA serà ètica.

CONCLUSIONS

Hem volgut contribuir a la discussió sobre la condició de la IA contestant a les tres preguntes clau: què és pròpiament la IA?, quin és l'estatus epistemològic de la IA?, i què és l'ètica de la IA? i que corresponen a un abordatge ontològic, epistemològic i ètic

Respecte a la primera pregunta hem clarificat que la IA és una disciplina científica que forma part del camp de la informàtica destinat al desenvolupament d'algoritmes que tenen com a objectiu dur a terme tasques concretes que normalment requeririen de la intel·ligència humana per a ser realitzades. Aquesta definició, identifica la IA en funció d'un objectiu: comportar-se de tal manera que si una persona es comportés així, diríem que és intel·ligent. L'objectiu es relaciona amb la funció que es pretén que realitzi i la finalitat per la qual ha estat fabricada. També es persegueix la finalitat que els sistemes d'IA realitzin les tasques que li encomanem sense la necessitat de supervisió constant (autonomia), delegant la presa de decisions en ells. Tanmateix, només poden fer les coses per a les quals han estat programats per fer. En aquest sentit, també podem parlar de la IA com una tecnologia que utilitza un conjunt de tècniques de diverses disciplines científiques tals com l'estadística, les matemàtiques, la neurociència, les ciències de la computació, etc. Un fet, aquest, que ens anuncia que la IA depèn del desenvolupament d'aquestes disciplines i que es deu a la confluència de tals disciplines en un objectiu comú. En definitiva, la IA és pròpiament el terreny on conflueixen aquelles disciplines destinades a produir artefactes d'acord amb aquestes finalitats.

Contestant a la segona pregunta, per analitzar l'estatus epistèmic de la IA (com funciona la IA i com accedeix al coneixement), hem plantejat que l'epistemologia de la IA es basa en un procés diferent del coneixement humà. Mentre que la intel·ligència humana s'estructura al voltant de l'experiència, la reflexió, i la vivència, la IA es fonamenta en el processament de grans quantitats de dades (*big data*) i l'extracció de patrons estadístics mitjançant algoritmes. El coneixement de la IA no es pot descriure com a comprensió o consciència en el sentit humà, sinó com una acumulació d'informació organitzada a partir d'*inputs* externs que permeten predir, categoritzar o executar tasques determinades amb certa autonomia.

Això implica que la IA no genera coneixement en termes de comprensió conceptual o intencionalitat, sinó que estableix correlacions entre conjunts de dades que l'ajuden a optimitzar el seu rendiment. Es tracta d'una aproximació epistèmica que resulta ser funcional, però també està limitada pel fet que la IA no té accés al significat de la informació que processa: el seu *coneixement* és operatiu, no és vivencial ni contextual. L'estatus epistèmic de la IA és

instrumental: no produeix coneixement, en tot cas estableix correlacions i fa prediccions que depenen de les dades amb les quals ha estat entrenada.

En vistes d'això, un aspecte problemàtic en el context de l'estatus epistèmic de la IA és el de les caixes negres: sistemes opacs que proporcionen resultats útils i que no podem conèixer com arriben a aquestes conclusions. Aquest és un aspecte que comporta una dificultat addicional: esperem delegar la nostra autonomia a sistemes que prenen decisions sense que nosaltres puguem traçar el recorregut de les seves deduccions. La opacitat del seu funcionament, com a mínim, hauria de generar incertesa sobre la fiabilitat de les conclusions que ens ofereix la IA i la confiança que hi podem dipositar. En darrer terme, l'estatus epistemològic de la IA es caracteritza per ser funcional i instrumental. No tenim raons per a pensar que sistemes basats en IA tinguin la capacitat de comprendre res del que fan, perquè no estan basats en el coneixement, sinó en l'extracció i predicció de patrons estadístics.

Contestant a la tercera pregunta sobre què és l'ètica de la IA, hem vist que es tracta d'un camp encara en desenvolupament que reflexiona sobre les implicacions ètiques i polítiques de la tecnologia. La tecnologia, inclosa la IA, no és neutral: està impregnada de valors i intencions dels seus desenvolupadors, i s'implementa en un context social, polític i cultural que no és mai axiològicament neutre. S'ha argumentat que aquest fet exigeix una ètica tecnològica capaç de garantir que la IA serveixi al benestar social i no només a interessos empresarials.

Si entenem l'ètica de la IA com quelcom semblant a incorporar algun tipus d'algoritme ètic als sistemes d'IA, s'ha justificar que l'ètica no es pot reduir a un algoritme que la modelitzi ni a un llenguatge logicoformal. A diferència dels sistemes d'IA, els éssers humans prenen decisions en funció de valors, contextos i sensibilitat, no simplement seguint regles lògiques. Els intents de dotar la IA de mecanismes basats en alguna cosa semblant a la reflexió ètica fracassen, perquè no poden capturar la complexitat de les decisions humanes ni la diversitat de valors, contextos i sensibilitats.

Quan deleguem tasques als sistemes d'IA es planteja l'ineludible problema de qui s'hauria de responsabilitzar davant de mals funcionaments o de les conseqüències negatives que comportin les decisions preses. Mentre que els sistemes tecnològics que ens assisteixen poden prendre algunes decisions que afecten a les persones, la responsabilitat continua essent humana per permetre-ho. Per tant, la responsabilitat ètica ha de romandre en mans humanes.

La falta de transparència dels sistemes que funcionen com a caixes negres suposa un dels grans reptes: hauríem de delegar tasques que tenen a veure amb la vida humana a sistemes que no sabem com funcionen? S'han presentat les dues condicions que s'han de donar per atribuir responsabilitat: que se sàpiga allò que s'està fent (condició epistèmica) i el control que se'n tingui (condició de control). S'ha al·legat que aquestes condicions haurien de ser factors restrictius en el desenvolupament de noves tecnologies, és a dir, que si no es pot assegurar el control i el coneixement d'una tecnologia basada en IA que ha de prendre decisions relacionades amb la vida, serà convenient no implementar-la. Finalment, concloem que encara que l'ètica de la IA sigui un terreny fèrtil des d'on orientar la reflexió ètica al voltant de la IA, convindria escapar d'una ètica única i específica sobre la IA que reflexioni de forma aïllada i abstracta respecte dels contextos, pràctiques i finalitats a què serveix.

REFERÈNCIES BIBLIOGRÀFIQUES

Agüera, B. (2022). Do Large Language Models Understand Us? *Daedalus*, *151*(2), 183-197. https://doi.org/10.1162/DAED_A_01909

AI Index Steering Committee. (2024). *Artificial Intelligence Index Report 2024*. https://aiindex.stanford.edu/wp-content/uploads/2024/05/HAI_AI-Index-Report-2024.pdf

Aristóteles. (2014). *Metafísica*. Gredos.

Aristòtil. (1995). *Física*. Editorial Gredos.

Aristòtil. (2015). *De l'ànima*. Bernat Metge.

Armstrong, J. (2002). *Conditions of Love: The Philosophy of Intimacy*. Penguin.

Baars, B. J. (1988). *A cognitive theory of consciousness*. Cambridge University Press.

Barbaras, Renaud. (1998). *Le tournant de l'expérience : recherches sur la philosophie de Merleau-Ponty*. Vrin.

Barclays. (2023). *A Guide to the New Age of AI*.

Barrow, J. D., Davies, P. C. W., & Jr, C. (2004). Science and Ultimate Reality: Quantum Theory, Cosmology, and Complexity. En *Science and Ultimate Reality*. Cambridge University Press. https://doi.org/10.1017/CBO9780511814990

Bastos, D., Fernández-Caballero, A., Pereira, A., & Rocha, N. P. (2022). Smart City Applications to Promote Citizen Participation in City Management and Governance: A Systematic Review. *Informatics 2022, Vol. 9, Page 89*, *9*(4), 89. https://doi.org/10.3390/INFORMATICS9040089

Bechtel, W. (2007). Mental mechanisms: Philosophical perspectives on cognitive neuroscience. En *Mental Mechanisms: Philosophical Perspectives on Cognitive Neuroscience* (1a ed.). Psychology Press. https://doi.org/10.4324/9780203810095/MENTAL-MECHANISMS-WILLIAM-BECHTEL/ACCESSIBILITY-INFORMATION

Bestmann, S., Baudewig, J., Siebner, H. R., Rothwell, J. C., & Frahm, J. (2004). Functional MRI of the immediate impact of transcranial magnetic stimulation on cortical and subcortical motor circuits. *The European journal of neuroscience*, *19*(7), 1950-1962. https://doi.org/10.1111/J.1460-9568.2004.03277.X

Blitz, D. (1992). Emergent Evolution: Qualitative Novelty and the Levels of Reality. En *Emergent Evolution*. Springer Netherlands. https://doi.org/10.1007/978-94-015-8042-7

Block, N. (1978). Troubles with functionalism. *Minnesota Studies in the Philosophy of Science*, *9*, 261-325.

Block, N. (1990). The Computer Model of the Mind. En D. N. Osherson & E. E. Smith (Ed.), *Thinking: An Invitation to Cognitive Science* (Vol. 3, p. 247-289). Mass: MIT Press.

Block, N. (1993). Holism, hyper-analyticity and hyper-compositionality. *Mind & Language*, *8*(1), 1-26. https://doi.org/10.1111/J.1468-0017.1993.TB00267.X

Block, N. (1996). Mental paint and mental latex. *Philosophical Issues*, *7*, 49. https://doi.org/10.2307/1522889

Boden, M. (2018). *Artificial Intelligence: A Very Short Introduction* (Vol. 1). Oxford University Press. https://doi.org/10.1093/ACTRADE/9780199602919.001.0001

Boden, M. A. (2016). *AI: Its nature and future* . OUP Oxford.

Boden, M., Bryson, J., Caldwell, D., Dautenhahn, K., Edwards, L., Kember, S., Newman, P., Parry, V., Pegman, G., Rodden, T., Sorrell, T., Wallis, M., Whitby, B., & Winfield, A. (2017). Principles of robotics: regulating robots in the real world. *Connection Science*, *29*(2), 124-129. https://doi.org/10.1080/09540091.2016.1271400

Bressler, S. L., Tang, W., Sylvester, C. M., Shulman, G. L., & Corbetta, M. (2008). Top-Down Control of Human Visual Cortex by Frontal and Parietal Cortex in Anticipatory Visual Spatial Attention. *The Journal of Neuroscience*, *28*(40), 10056-10061. https://doi.org/10.1523/JNEUROSCI.1776-08.2008

Bringsjord, S., Bello, P., & Ferrucci, D. (2000). Creativity, the Turing test, and the (better) Lovelace Test. *Minds and Machines*, *11*(1), 3-27. https://doi.org/10.1023/A:1011206622741

Bueno, G. (1996). El sentido de la Vida: seis lecturas de la filosofía moral. En *EL Sentido de la Vida. Pentalfa Ediciones*. Pentalfa Ediciones.

Bueno, G. (2006). Sobre los derechos de los simios. En *Zapatero y el Pensamiento Alicia* (p. 109-158). Temas de Hoy.

Bunge, M. (1977). Towards a Technoethics. *Monist*, *60*(1), 96-107. https://doi.org/10.5840/MONIST197760134

Bunge, M. (2015). *Materia y mente: una investigación filosófica* (1ª). Laetoli.

Burnyeat, M. (1995). Is Aristotle's Philosophy of Mind Still Credible? Essays on Aristotle's De Anima. Eds. Martha Nussbaum and Amélie Rorty. *Clarendon Press*, 15-26. https://doi.org/10.1007/978-3-319-04361-6_3

Butlin, P. (2022). Machine Learning, Functions and Goals. *Croatian Journal of Philosophy*, *22*(66), 351-370. https://doi.org/10.52685/CJP.22.66.5

Butlin, P., Long, R., Elmoznino, E., Bengio, Y., Birch, J., Constant, A., Deane, G., Fleming, S. M., Frith, C., Ji, X., Kanai, R., Klein, C., Lindsay, G., Michel, M., Mudrik, L., Peters, M. A. K., Schwitzgebel, E., Simon, J., & VanRullen, R. (2023). *Consciousness in Artificial Intelligence: Insights from the Science of Consciousness*. https://doi.org/https://doi.org/10.48550/arXiv.2308.08708

Camps, V. (2016). *Elogio de la duda* (7a ed.). Arpa.

Caston, V. (2002). Aristotle on consciousness. *Mind*, *111*(444), 751-815. https://doi.org/10.1093/MIND/111.444.751

Caston, V. (2005). The Spirit and the Letter: Aristotle on Perception. Themes from the Work of Richard Sorabji. Ed. Ricardo Salle. *Oxford University Press*, 245-320.

Chaitin, G. (2004). Meta Math! The Quest for Omega. *Choice Reviews Online*, *43*(07), 43-4073-43-4073. https://doi.org/10.5860/choice.43-4073

Chalmers, D. J. (2003). Consciousness and its Place in Nature. En S. Stich & T. Warfield (Ed.), *Blackwell Guide to Philosophy of Mind* (p. 102-143). Blackwell.

Cheng, J., Novati, G., Pan, J., Bycroft, C., Žemgulyte, A., Applebaum, T., Pritzel, A., Wong, L. H., Zielinski, M., Sargeant, T., Schneider, R. G., Senior, A. W., Jumper, J., Hassabis, D., Kohli, P., & Avsec, Ž. (2023). Accurate proteome-wide missense variant effect prediction with AlphaMissense. *Science*, *381*(6664). https://doi.org/10.1126/SCIENCE.ADG7492/SUPPL_FILE/SCIENCE.ADG7492_DATA_S1_TO_S9.ZIP

Churchland, P. (1990). Could a Machine Think? *Scientific American*, *262*(1), 32-39.

Churchland, P. M. (1981). Eliminative Materialism and the Propositional Attitudes. *The Journal of Philosophy*, *78*, 67-90. https://doi.org/10.2307/2026571

Clark, A. (2011). *Supersizing the mind: embodiment, action, and cognitive extension*. Oxford University Press.

Clark, A., & Chalmers, D. (1998). The Extended Mind . *Alanysis*, *58*(1), 7-19.

Coeckelbergh, M. (2011). Humans, Animals, and Robots. *International Journal of Social Robotics*, *3*(2), 197-204. https://doi.org/10.1007/S12369-010-0075-6

Coeckelbergh, M. (2021). *Ética de la inteligencia artifical*. Cátedra.

Coeckelbergh, M. (2023). *La filosofía política de la inteligencia artificial*. Cátedra.

Coeckelbergh, M. (2024). *La ética de los robots* (1a ed.). Cátedra.

Colbert, D., Malone, A., Barrett, S., & Roche, B. (2020). The Relational Abilities Index+: Initial Validation of a Functionally Understood Proxy Measure for Intelligence. *Perspectives on Behavior Science*, *43*(1), 189-213. https://doi.org/10.1007/S40614-019-00197-Z/TABLES/3

Copeland, B. J. (2001). The Turing Test . *Minds and Machines*, *10*, 519-539.

Cortese, J. F. N. B., Cozman, F. G., Lucca-Silveira, M. P., & Bechara, A. F. (2022). Should explainability be a fifth ethical principle in AI ethics? *AI and Ethics*, *3*(1), 123-134. https://doi.org/10.1007/S43681-022-00152-W

Crawford, K. (2021). *Atlas of AI: Power, Politics, and the Planetary Costs of Artificial Intelligence*. Yale University Press.

Crawford, K., & Paglen, T. (2021). Excavating AI: the politics of images in machine learning training sets. *AI and Society*. https://doi.org/10.1007/S00146-021-01162-8

Damasio, A. (2009). *En busca de Spinoza. Neurobiología de la emoción y los sentimientos* (6ª). Crítica.

Davidson, D. (1980). Mental Events. En *Essays on Actions and Events* (p. 207-225). Clarendon Press.

Davies, P. C. W. (2004). John Archibald Wheeler and the clash of ideas. *Cambridge University Press* , 3-23.

de Fine Licht, K., & de Fine Licht, J. (2020). Artificial intelligence, transparency, and public decision-making. *AI & SOCIETY*, *35*(4), 917-926. https://doi.org/10.1007/s00146-020-00960-w

Degli-Esposti, S. (2023). *La ética de la inteligencia artificial*. CSIC.

Dehaene, S., Kerszberg, M., & Changeux, J. P. (1998). A neuronal model of a global workspace in effortful cognitive tasks. *Proceedings of the National Academy of Sciences of the United States of America*, *95*(24), 14529-14534. https://doi.org/10.1073/PNAS.95.24.14529

Dehaene, S., & Naccache, L. (2001). Towards a cognitive neuroscience of consciousness: basic evidence and a workspace framework. *Cognition*, *79*(1-2), 1-37. https://doi.org/10.1016/S0010-0277(00)00123-2

Dehghani, M., Forbus, K., Tomai, E., & Klenk, M. (2011). An integrated reasoning approach to moral decision making. En Anderson M & Anderson S L (Ed.), *Machine Ethics* (p. 422-441). Cambridge University Press. https://doi.org/10.1017/CBO9780511978036.024

Dennet, D. C. (2017). *DE LAS BACTERIAS A BACH La evolución de la mente*. Pasado & Presente.

Dennett, D. C. (1989). *The Intentional Stance* . MIT Press.

Descartes, R. (1987). *Carta del autor a quien tradujo: Los principios de la filosofía : Vol. Cuaderno 44*. Universidad Nacional Autónoma de México.

Descartes, R. (2010). *El Discurso del Método*. Austral - Espasa Calpe.

Diéguez, A. (2017). *Transhumanismo*. Herder.

Doshi-Velez, F., & Kim, B. (2017). Towards A Rigorous Science of Interpretable Machine Learning. *Machine Learning*.

Double, R. (1999). *Beginning philosophy*. Oxford University Press.

Dreyfus, H. L. . (1994). *What computers still can't do: a critique of artificial reason*. MIT Press.

Du Sautoy, M. (2020). *Programados para crear*. Acantilado.

Echeverría, J. (2010). Tecnociencia, tecnoética y tecnoaxiología. *Revista Colombiana de Bioética*, *5*(1), 142-152.

Elish, M. C. (2019). Moral Crumple Zones: Cautionary Tales in Human-Robot Interaction. *Engaging Science, Technology, and Society*, *5*, 40-60. https://doi.org/10.17351/ests2019.260

Etxeberria, A., & Casado, A. (2008). Autonomía, vida y bioética. *Ludus Vitalis*, *17*(30), 213-216.

European Comission. (2024). *AI Pact What is the AI Pact?* https://ec.europa.eu/eusurvey/runner/68fd7335-f477-b1a7-f52f-e51b60a825b5

European Commission. (2021). *Proposal for a Regulation of the European Parliament and of the Council laying down harmonised rules on Artificial Intelligence (Artificial Intelligence Act) and amending certain union legislative acts COM/2021/206*. https://eur-lex.europa.eu/legal-content/EN/ALL/?uri=CELLAR:e0649735-a372-11eb-9585-01aa75ed71a1

European Parliament. (2024). *European Parliament legislative resolution of 13 March 2024 on the proposal for a regulation of the European Parliament and of the Council on laying down harmonised rules on Artificial Intelligence (Artificial Intelligence Act) and amending certain Union Legislative Acts (COM(2021)0206-C9-0146/2021-2021/0106(COD))*. https://www.europarl.europa.eu/doceo/document/TA-9-2024-0138_EN.pdf

Ferraris, M. (2022). Realismo Transcendental. *Disputatio. Philosophical Research Bulletin*, *11*(20), 145-158.

Fischer, J. M., & Ravizza, M. (1998). *Responsibility and Control: A Theory of Moral Responsibility*. Cambridge University Press.

Flores, A. (2011). Development of Computational Thinking in Discrete Mathematics Training . *Lámpsakos*, *5*, 28-33.

Floridi, L. (2013). Distributed morality in an information society. *Science and Engineering Ethics*, *19*(3), 727-743. https://doi.org/10.1007/S11948-012-9413-4

Floridi, L. (2015). The Onlife Manifesto. *The Onlife Manifesto: Being Human in a Hyperconnected Era*, 7-13. https://doi.org/10.1007/978-3-319-04093-6_2

Floridi, L. (2022). Ultraintelligent Machines, Singularity, and Other Sci-fi Distractions about AI. *Lavoro Diritti Europa*, *3*, 1-19.

Floridi, L., & Cowls, J. (2019). A Unified Framework of Five Principles for AI in Society. *Harvard Data Science Review*, *1*(1). https://doi.org/10.1162/99608F92.8CD550D1

Floridi, L., Cowls, J., Beltrametti, M., Chatila, R., Chazerand, P., Dignum, V., Luetge, C., Madelin, R., Pagallo, U., Rossi, F., Schafer, B., Valcke, P., & Vayena, E. (2018). AI4People-An Ethical Framework for a Good AI Society: Opportunities, Risks, Principles, and Recommendations. *Minds and machines*, *28*(4), 689-707. https://doi.org/10.1007/S11023-018-9482-5

Floridi, L., & Sanders, J. W. (2004). On the morality of artificial agents. *Minds and Machines*, *14*(3), 349-379. https://doi.org/10.1023/B:MIND.0000035461.63578.9D/METRICS

Flower, T. P., Gribble, M., & Ridley, A. R. (2014). Deception by flexible alarm mimicry in an african bird. *Science*, *344*(6183), 513-516. https://doi.org/10.1126/SCIENCE.1249723/SUPPL_FILE/FLOWER.SM.PDF

Fodor, J. A. (1984). *El lenguaje del pensamiento* . Alianza.

Fodor, J. A. (1985). Fodor's Guide to Mental Representation: The Intelligent Auntie's Vade-Mecum. *Mind*, *94*(373), 76-100. https://doi.org/10.1093/MIND/XCIV.373.76

Fowler, D., & Robson, E. (1998). Square Root Approximations in Old Babylonian Mathematics: YBC 7289 in Context. *Historia Mathematica*, *25*(4), 366-378. https://doi.org/10.1006/HMAT.1998.2209

Funk, M., & Coeckelbergh, M. (2020). (Technical) Autonomy as Concept in Robot Ethics. *Springer Nature Switzerland*, *25*, 59-65. https://doi.org/10.1007/978-3-030-24074-5_12

Future of Life Institute. (2023, març 22). *Pause Giant AI Experiments: An Open Letter*.

Gabriel, M. (2016). *Yo No Soy Mi Cerebro: Filosofía de la mente para el siglo XXI* (2a ed.). Pasado y Presente.

Gabriel, M. (2017). *Sentido y existencia. Una ontología realista*. Herder.

Gabriel, M. (2019). *El sentido del pensamiento*. Pasado y presente.

Gardner, H. (2019). *Inteligencias múltiples*. Planeta.

Génova, G. (2023). Inteligencia artificial: explicabilidad, racionalidad y responsabilidad profesional del ingeniero. *El Basilisco*, *59*, 44-48.

Geschwind, N. (1970). The organization of language and the brain. *Science*, *170*(3961), 940-944. https://doi.org/10.1126/SCIENCE.170.3961.940

Gibson, J. J. (1972). A Theory of Direct Visual Perception. In J. Royce, W. Rozenboom (Eds). The Psychology of Knowing. En *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science* (Número 5). Gordon & Breach.

Goldstein, E. B. (2006). *Sensación y percepción* (6a ed.). Ediciones Paraninfo.

González, R., & Vergauwen, R. (2005). On the versimilitude of Artificial Intelligence. *Logique et Analyse*, *189*(192), 323-350.

Goodfellow, I. J., Shlens, J., & Szegedy, C. (2014). Explaining and Harnessing Adversarial Examples. *3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings*.

Gordon, J.-S. (2020). Building Moral Robots: Ethical Pitfalls and Challenges. *Science and engineering ethics*, *26*(1), 141-157. https://doi.org/10.1007/S11948-019-00084-5

Gruber, T. R. (1993). A translation approach to portable ontology specifications. *Knowledge Acquisition*, *5*(2), 199-220. https://doi.org/10.1006/KNAC.1993.1008

Guarini, M. (2006). Particularism and the classification and Classification of moral cases. *IEEE Intelligent Systems*, *21*(4), 22-28. https://doi.org/10.1109/MIS.2006.76

Hacking, I. (2000). *The social construction of what?* Harvard University Press.

Haidt, J. (2001). The emotional dog and its rational tail: A social intuitionist approach to moral judgment. *Psychological Review*, *108*(4), 814-834. https://doi.org/10.1037/0033-295X.108.4.814

Hare, R. D. (1973). Orienting and Defensive Responses to Visual Stimuli. *Psychophysiology*, *10*(5), 453-464. https://doi.org/10.1111/J.1469-8986.1973.TB00532.X

Heil, J. (2000). *Philosophy of Mind: A Contemporary Introduction* . Routledge.

Hume, D. (1980). *Investigación sobre el conocimiento humano*. Alianza.

Husserl, E. (1980). *Experiencia y Juicio. Investigaciones acerca de la genealogía de la lógica*. UNAM.

Iglesias, J. E., Billot, B., Balbastre, Y., Magdamo, C., Arnold, S. E., Das, S., Edlow, B. L., Alexander, D. C., Golland, P., & Fischl, B. (2023). SynthSR: A public AI tool to turn heterogeneous clinical brain scans into high-resolution T1-weighted images for 3D morphometry. *Science Advances*, *9*(5). https://doi.org/10.1126/SCIADV.ADD3607

Irvine, E. (2013). Measures of Consciousness. *Philosophy Compass*, *8*(3), 285-297. https://doi.org/10.1111/PHC3.12016

Jeans, J. (1937). *The mysterious universe*. Penguin Books.

Ji, J., Qiu, T., Chen, B., Zhang, B., Lou, H., Wang, K., Duan, Y., He, Z., Zhou, J., Zhang, Z., Zeng, F., Dai, J., Pan, X., Yee Ng Aidan, K. O., Xu, H., Tse Jie Fu, B., McAleer, S., Yang, Y., Wang, Y., … Gao, W. (2024). AI Alignment: A Comprehensive Survey. *Arxiv*, 1-102.

Jonas, H. (1979). Toward a Philosophy of Technology. *The Hastings Center Report*, *9*(1), 43. https://doi.org/10.2307/3561700

Jonas, H. (1995). *El principio de responsabilidad. Ensayo de una ética para la civilización tecnológica*. Herder.

Jones, S. E. (2006). *Agains Technology. From the Luddites to Neo-Luddism*. Taylor & Francis Group.

Kaminski, J., Waller, B. M., Diogo, R., Hartstone-Rose, A., & Burrows, A. M. (2019). Evolution of facial muscle anatomy in dogs. *Proceedings of the National Academy of Sciences of the United States of America*, *116*(29), 14677-14681. https://doi.org/10.1073/PNAS.1820653116

Kaplan, J. (2017). *Inteligencia artifical: Lo que todo el mundo debe saber*. TEELL.

Kastrup, B. (2014). *Why Materialism Is Baloney: How True Skeptics Know There is no Death and Fathom Answers to Life, the Universe and Everything*. Iff Books.

Kaufman, D. (2014). Cartesian Substances, Individual Bodies, and Corruptibility. *Res Philosophica*, *91*(1), 71-102. https://doi.org/10.11612/resphil.2014.91.1.4

Kurzweil, R. (2019). *Cómo crear una mente: el secreto del pensamiento humano*. Lola Books.

Lamme, V. A. F. (2020). Visual Functions Generating Conscious Seeing. *Frontiers in Psychology*, *11*(83). https://doi.org/10.3389/FPSYG.2020.00083/FULL

Latorre Sentís, J. I. (2019). *Ética para máquinas*. Ariel.

Latour, B. (1996). Social theory and the study of computerized work sites. En W. J. Orlinokowski & G. Walsham (Ed.), *Information Technology and Changes in Organizational Work* (p. 295-307). Chapman and Hall.

Legg, S., & Hutter, M. (2007). Universal Intelligence: A Definition of Machine Intelligence. *Minds and Machines*, *17*(4).

Leibniz, G. W. (1951). The Art of Discovery. En Philip P. Wiener (Ed.), *Leibniz: Selections*. Scribner. https://doi.org/10.1007/978-3-319-17912-4

Leondes, C. T. (2001). *Expert Systems: The Technology of Knowledge Management and Decision Making for the 21st Century* (1a ed.). Academic Press.

Lin, P., Abney, K., & Bekey, G. A. (Ed.). (2014). *Robot Ethics: The Ethical and Social Implications of Robotics. Intelligent Robotics and Autonomous Agents*. MIT Press.

Loi, M., Heitz, C., & Christen, M. (2020). A Comparative Assessment and Synthesis of Twenty Ethics Codes on AI and Big Data. *2020 7th Swiss Conference on Data Science*, 41-46. https://doi.org/10.1109/SDS49233.2020.00015

López de Mántaras, R. (2018). El futura de la IA: hacia inteligencias artificiales realmente inteligentes. En *¿Hacia una nueva Ilustración? Una década trascendente* (p. 160-174). BBVA.

Lowder, M. W., & Gordon, P. C. (2015). Natural forces as agents: reconceptualizing the animate-inanimate distinction. *Cognition*, *136*, 85-90. https://doi.org/10.1016/J.COGNITION.2014.11.021

Lu, H., Diaz, D. J., Czarnecki, N. J., Zhu, C., Kim, W., Shroff, R., Acosta, D. J., Alexander, B. R., Cole, H. O., Zhang, Y., Lynd, N. A., Ellington, A. D., & Alper, H. S. (2022). Machine learning-aided engineering of hydrolases for PET depolymerization. *Nature 2022 604:7907*, *604*(7907), 662-667. https://doi.org/10.1038/s41586-022-04599-z

Mach, E. (1976). Knowledge and Error. En *Knowledge and Error*. Reidel. https://doi.org/10.1007/978-94-010-1428-1

Madhavan, P., & Wiegmann, D. A. (2007). Effects of information source, pedigree, and reliability on operator interaction with decision support systems. *Human factors*, *49*(5), 773-785. https://doi.org/10.1518/001872007X230154

Malach, R. (2022). The Role of the Prefrontal Cortex in Conscious Perception: The Localist Perspective. *Journal of Consciousness Studies*, *29*(7-8), 93-114. https://doi.org/10.53765/20512201.29.7.093

Marchese, D. (2023). How Do We Ensure an A.I. Future That Allows for Human Thriving? . *The New York Times Magazine*.

Marcus, G., & Davis, E. (2020). GPT-3, Bloviator: OpenAI's language generator has no idea what it's talking about. *MIT Technology Review*.

Mashour, G. A., Roelfsema, P., Changeux, J. P., & Dehaene, S. (2020). Conscious Processing and the Global Neuronal Workspace Hypothesis. *Neuron*, *105*(5), 776-798. https://doi.org/10.1016/J.NEURON.2020.01.026

Matthias, A. (2004). The responsibility gap: Ascribing responsibility for the actions of learning automata. *Ethics and Information Technology*, *6*(3), 175-183. https://doi.org/10.1007/S10676-004-3422-1

McCarthy, J., Minsky, M. L., Rochester, N., & Shannon, C. E. (1955). A Proposal for the Dartmouth Summer Research Project on Artificial Intelligence, August 31, 1955. *AI Magazine*, *27*(4), 12-14. https://doi.org/10.1609/AIMAG.V27I4.1904

McCarthy, J., Minsky, M. L., Rochester, N., & Shannon, C. E. (2006). A Proposal for the Dartmouth Summer Research Project on Artificial Intelligence, August 31, 1955. *AI Magazine*, *27*(4), 12-12. https://doi.org/10.1609/AIMAG.V27I4.1904

McCorduck, P. (1979). Machines Who Think: A Personal Inquiry into the History and Prospects of Artificial Intelligence. *Choice Reviews Online*, *42*(02), 42-0916-42-0916. https://doi.org/10.5860/CHOICE.42-0916

McCulloch, W. S., & Pitts, W. (1947). How we know universals the perception of auditory and visual forms. *The Bulletin of Mathematical Biophysics*, *9*(3), 127-147. https://doi.org/10.1007/BF02478291/METRICS

McGinn, C. (1991). The Problem of Consciousness: Essays Toward a Resolution. En *The Philosophical Review* (Vol. 102, Número 2). Blackwell. https://doi.org/10.2307/2186044

McGinn, C. (1993). Apes, Humans, Aliens, Vampires and Robots . En P. Cavalieri & P. Singer (Ed.), *The Great Ape Project* (p. 146-151). St. Martin's Griffin.

Merleau-Ponty, M. (1986). *El Ojo y El Espiritu*. Ediciones Paidos.

Merleau-Ponty, M. (1993). Fenomenología de la Percepción. En *Editorial Alta ya*. Planeta Agostini.

Merleau-Ponty, Maurice. (2009). *Elogio y posibilidad de la filosofía*. Editorial Universidad de Almería.

Metzinger, T. (2003). Being No One: The Self-Model Theory of Subjectivity. En *Being No One*. The MIT Press. https://doi.org/10.7551/MITPRESS/1551.001.0001

Metzinger, T. (2018). *El túnel del yo. Ciencia de la mente y mito del sujeto*. Enclave de libros.

Metzinger, T. (2021). Artificial Suffering: An Argument for a Global Moratorium on Synthetic Phenomenology. *Journal of Artificial Intelligence and Consciouness*, *8*(1), 43-66. https://doi.org/10.1142/S270507852150003X

Michel, M., & Doerig, A. (2021). A new empirical challenge for local theories of consciousness. *Mind and Language*, *37*(5), 840-855. https://doi.org/10.1111/MILA.12319

Michel, M., & Lau, H. (2021). Higher-order theories do just fine. *Cognitive neuroscience*, *12*(2), 77-78. https://doi.org/10.1080/17588928.2020.1839402

Moore, A. (2018). *The infinite*. Routledge.

Moulines, C. U. (1991). *Pluralidad y recursión. Estudios epistemológicos* . Alianza Editorial.

Moya, C. J. (2006). *Filosofía de la mente* (2a ed.). PUV.

Müller, V. C. (2020). Ethics of Artificial Intelligence and Robotics. *Stanford Encyclopedia of Philosophy*, *7*(3), 1-70. https://doi.org/10.29012/JPC.V7I3.404

Nagel, T. (1974). Philosophical Review What Is It Like to Be a Bat? *Philosophical Review*, *83*(4), 435-450.

Nagel, T. (2002). *Concealment and exposure: and other essays*. Oxford University Press.

Nemitz, P. (2018). Constitutional democracy and technology in the age of artificial intelligence. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, *376*(2133). https://doi.org/10.1098/RSTA.2018.0089

Newell, A., & Simon, H. A. (1976). Computer science as empirical inquiry. *Communications of the ACM*, *19*(3), 113-126. https://doi.org/10.1145/360018.360022

O'Connor, T. (1994). Emergent properties. *American Philosophical Quarterly*, *31*(2), 91-104.

O'neil, C. (2018). *Armas de destrucción matemática: como el big data aumenta la desigualdad y la amenaza democrática*. Capital Swing Libros.

Papineau, D. (2002). Thinking about Consciousness. En *Thinking about Consciousness*. Oxford University Press. https://doi.org/10.1093/0199243824.001.0001

Paradis, M. (2004). *A Neurolinguistic Theory of Bilingualism*. John Benjamins Publishing Company. https://doi.org/10.1075/SIBIL.18

Pareto Boada, J. (2021). Prolegómenos a una ética para la robótica social. *Dilemata*, *34*, 71-87.

Pareto, J., & Torras, C. (2024). To each Technology Its Own Ethics? A Reply to Sætra & Danaher (and Their Critics). *Philosophy and Technology*, *37*(3), 1-6. https://doi.org/10.1007/S13347-024-00798-W/METRICS

Pasquinelli, M., & Joler, V. (2021). The Nooscope manifested: AI as instrument of knowledge extractivism. *AI and Society*, *36*(4), 1263-1280. https://doi.org/10.1007/S00146-020-01097-6/METRICS

Pia, L., Neppi-Modona, M., Ricci, R., & Berti, A. (2004). The anatomy of anosognosia for hemiplegia: a meta-analysis. *Cortex*, *40*(2), 367-377. https://doi.org/10.1016/S0010-9452(08)70131-X

Picard, R. W. (1997). *Affective Computing*. The MIT Press. https://doi.org/10.7551/MITPRESS/1140.001.0001

Pierce, C. S. (1976). *The New Elements of Mathematics* (Eisele Carolyn, Ed.; Vol. 4). The Hague - Mouton.

Pomerleau, D. A. (1991). Efficient Training of Artificial Neural Networks for Autonomous Navigation. *Neural Computation*, *3*(1), 88-97.

Popper, K. R. (1974). *Conocimiento objetivo. Un enfoque evolucionista*. Tecnos.

Potter, V. R. (1970). Bioethics, the Science of Survival. *Perspectives in Biology and Medicine*, *14*(1), 127-153. https://doi.org/10.1353/PBM.1970.0015

Povinelli, D. J. (2001). The Self: Elevated in consciouness and extended in time. En C. Moore & K. Lemmon (Ed.), *The Self in Time: Developmental perspectives* (p. 75-95). Lawrence Erlbaum Associates Publishers. https://doi.org/10.4324/9781410600684

Purves, D., Brannon, E., Cabeza, R., Huettel, S. A., & LaBar, K. S. (2008). Principles of Cognitive Neuroscience. En *Principles of Cognitive Neuroscience*. Sinauer Associates.

Putnam, H. (1975). The nature of mental states. En *Philosophical Papers* (p. 429-440). Cambridge University Press. https://doi.org/10.1017/CBO9780511625251.023

Putnam, H. (1988). *Representation and reality*. MIT Press.

Quine, W. V. (2013). *Word and Object*. The MIT Press.

Ramírez, M. T. (2014). *La filosofía del quiasmo. Introducción al pensamiento de Maurice Merleau-Ponty*. Fondo de Cultura Económica.

Richardson, K. (2016). Sex Robot Matters: Slavery, the Prostituted, and the Rights of Machines. *IEEE Technology and Society Magazine*, *35*(2), 46-53. https://doi.org/10.1109/MTS.2016.2554421

Riesch, H. (2010). Simple or simplistic? Scientists' views on Occam's razor. *Theoria-Revista De Teoria Historia Y Fundamentos De La Ciencia*, *25*(1), 75-90. https://doi.org/10.1387/theoria.489

Robbins, S. (2019). A Misdirected Principle with a Catch: Explicability for AI. *Minds and Machines*, *29*(4), 495-514. https://doi.org/10.1007/S11023-019-09509-3/METRICS

Rochat, P. (2003). Five levels of self-awareness as they unfold early in life. *Consciousness and Cognition: An International Journal*, *12*(4), 717-731. https://doi.org/10.1016/S1053-8100(03)00081-3

Roediger, H., & McDermott, K. (1993). Implicit memory in normal human subject. *Handbook of Neuropsychology*, *8*.

Román, B. (2016). *Ética de los servicios sociales*. Herder. https://herdereditorial.com/etica-de-los-servicios-sociales-9788425437878?srsltid=AfmBOoqNtGl7uzmgRo_BbKs6QLdkh8KKkDhkaMFZsaboC1jzxp5WM9NN

Ronchi, A. M. . (2009). *eCulture : cultural content in the digital age* (1a ed.). Springer.

Rorty, R. (1983). *La filosofía y el espejo de la naturaleza*. Cátedra.

Rosa, H. (2021). *Lo indisponible*. Herder Editorial.

Rosenberg, A. (1981). *Sociobiology and the Preemption of Social Science* . Oxford Basil Blackwell.

Rosenblatt, F., Stieber, A., & Shatz, R. H. (1957). *The Perceptron. A perceiving and recognizing automaton (Project PARA)*.

Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence 2019 1:5*, *1*(5), 206-215. https://doi.org/10.1038/s42256-019-0048-x

Rump, K. M., Giovannelli, J. L., Minshew, N. J., & Strauss, M. S. (2009). The Development of Emotion Recognition in Individuals with Autism. *Child development*, *80*(5), 1447. https://doi.org/10.1111/J.1467-8624.2009.01343.X

Russell, B. (1982). *La evolución de mi pensamiento filosófico*. Alianza Editorial.

Russell, B. (1992). *Los problemas de la filosofía* . Labor.

Russell, B. (2016). *Fundamentos de filosofía*. DEBOLS!LLO.

Russell, S., & Norving, P. (2010). *Artificial Intelligence A Modern Approach Prentice-Hall* (3a ed.). Prentice-Hall.

Ryan, M. (2020). In AI We Trust: Ethics, Artificial Intelligence, and Reliability. *Science and engineering ethics*, *26*(5), 2749-2767. https://doi.org/10.1007/S11948-020-00228-Y

Sætra, H. S., & Danaher, J. (2022). To Each Technology Its Own Ethics: The Problem of Ethical Proliferation. *Philosophy and Technology*, *35*(4), 1-26. https://doi.org/10.1007/S13347-022-00591-7/FIGURES/2

Schmaltz, T. M. (2019). The Metaphysics of the Material World: Suárez, Descartes, Spinoza. *Oxford Academic*, 144-182. https://doi.org/10.1093/OSO/9780190070229.003.0005

Schrödinger, E. (2016). *Mente y materia*. Tusquets Editores.

Schwitzgebel, E., & Garza, M. (2020). Designing AI with Rights, Consciousness, Self-Respect, and Freedom. En L. S. Matthew (Ed.), *Ethics of Artificial Intelligence* (p. 459-479). Oxford University Press. https://doi.org/10.1093/OSO/9780190905033.003.0017

Searle, J. R. (1980). Minds, brains, and programs. *Behavioral and Brain Sciences*, *3*(3), 417-424. https://doi.org/10.1017/S0140525X00005756

Searle, J. R. (1985). *Mentes, cerebros y ciencia*. Catedra.

Searle, J. R. (2000). *El misterio de la conciencia*. Editorial Paidós.

Shulman, C., & Bostrom, N. (2021). Sharing the World with Digital Minds . En S. Clarke, H. Zohny, & J. Savulescu (Ed.), *Rethinking Moral Status*. Oxford University Press.

Silverberg, A. (2003). Psychological laws. *Erkenntnis*, *58*(3), 275-302. https://doi.org/10.1023/A:1022666700490

Singer, P. (1987). Animal Liberation or Animal Rights? *The Monist*, *70*(1), 3-14.

Sisko, J. E. (1996). Material Alteration and Cognitive Activity in Aristotle'ls De Anima. *Phronesis*, *41*(2), 138-157.

Skinner, B. F. (1985). Cognitive science and behaviourism. *British journal of psychology (London, England : 1953)*, *76 (Pt 3)*(3), 291-301. https://doi.org/10.1111/J.2044-8295.1985.TB01953.X

Smith, A. M. (2015). Responsibility as Answerability. *Inquiry*, *58*(2), 99-126. https://doi.org/10.1080/0020174X.2015.986851

Sobel, N., Prabhakaran, V., Desmond, J. E., Glover, G. H., Goode, R. L., Sullivan, E. V., & Gabriell, J. D. E. (1998). Sniffing and smelling: separate subsystems in the human olfactory cortex. *Nature 1998 392:6673*, *392*(6673), 282-286. https://doi.org/10.1038/32654

Sorabji, R. (1992). *Intentionality and Physiological Processes: Aristotle's Theory of Sense-Perception. Essays on Aristotle's De Anima. Eds. Martha Nussbaum and Amélie Rorty* (p. 195-225). Oxford University Press.

Stein-Perlman, Z., Weinstein-Raun, B., & Grace, K. (2022). Expert Survey on Progress in AI. *AI Impacts*, *3*.

Strachey, C. (1965). An impossible program. *The Computer Journal*, *7*(4), 313-313. https://doi.org/10.1093/COMJNL/7.4.313

Stubenberg, L. (2008). Neutral Monism: A Miraculous, Incoherent, and Mislabeled Doctrine?, Reduction and Elimination in the Philosophy of Science. *Contributions of the Austrian Ludwig Wittgenstein Society XVI*, 337-339.

Swaab, D. F. . (2014). *Somos nuestro cerebro: cómo pensamos, sufrimos y amamos* (6a ed.). Plataforma Editorial.

Swade, D. (2000). *The Difference Engine: Charles Babbage and the Quest to Build the First Computer*. Penguin.

Tallis, R. (2012). Aping Mankind: Neuromania, Darwinitis and the Misrepresentation of Humanity. En *TLS - The Times Literary Supplement* (1a ed., Número 5684). Routledge. https://doi.org/10.4324/9781315711386/APING-MANKIND-RAYMOND-TALLIS/ACCESSIBILITY-INFORMATION

Tegmark, M. (2023). The «Don't Look Up» Thinking That Could Doom Us With AI. *Time*.

Tegmark, Max. (2015). *Nuestro universo matemático : en busca de la naturaleza última de la realidad*. Antoni Bosch.

Thadani, N. N., Gurev, S., Notin, P., Youssef, N., Rollins, N. J., Ritter, D., Sander, C., Gal, Y., & Marks, D. S. (2023). Learning from prepandemic data to forecast viral escape. *Nature 2023 622:7984*, *622*(7984), 818-825. https://doi.org/10.1038/s41586-023-06617-0

Thompson, R. F. (2005). In search of memory traces. *Annual Review of Psychology*, *56*, 1-23. https://doi.org/10.1146/ANNUREV.PSYCH.56.091103.070239/CITE/REFWORKS

Tiwari, T., Tiwari, T., & Tiwari, S. (2018). How Artificial Intelligence, Machine Learning and Deep Learning are Radically Different? *International Journal of Advanced Research in Computer Science and Software Engineering*, *8*(2), 1. https://doi.org/10.23956/IJARCSSE.V8I2.569

Tomasik, B. (2014). Do Artificial Reinforcement-Learning Agents Matter Morally? *Foundational Research Institute*, 1-37.

Torralba Roselló, Francesc. (2022). *L'ètica algorítmica*. Edicions 62.

Trafton, J. G., McCurry, J. M., Zish, K., & Frazier, C. R. (2024). The Perception of Agency. *ACM Transactions on Human-Robot Interaction*, *13*(1), 23. https://doi.org/10.1145/3640011/ASSET/644520FE-F0F0-4AEC-B036-C6218204090B/ASSETS/GRAPHIC/THRI-2023-0027-F01.JPG

Tulio Ribeiro, M., Singh, S., Guestrin, C., Tulio Ribeiro, M., Singh, S., & Guestrin, C. (2016). «Why Should I Trust You?»: Explaining the Predictions of Any Classifier. *Psicholohycal Review*, *108*(4), 814-834. https://doi.org/10.48550/ARXIV.1602.04938

Turing, A. M. (1947). Lecture to the London Mathematical Society on 20 February 1947. En *The Charles Babbage Institute Reprint Series for the History of Computing* (Vol. 10).

Turing, A. M. (1950). Computing Machinery and Intelligence. *Source: Mind, New Series*, *59*(236), 433-460.

Turk, M. (2000). Perceptive Media: Machine Perception and Human Computer Interaction. *Chinese Journal of Computers*, *23*(12), 1235-1244.

Valdivieso, G., & Macedi, L. (2018). Neurociencias y psicoterapia: mecanismo top- down y bottom-up. *Rev Neuropsiquiatr*, 183-195. http://www.scielo.org.pe/pdf/rnp/v81n3/a06v81n3.pdf

van de Poel, I., Fahlquist, J. N., Doorn, N., Zwart, S., & Royakkers, L. (2012). The Problem of Many Hands: Climate Change as an Example. *Science and Engineering Ethics*, *18*(1), 49-67. https://doi.org/10.1007/S11948-011-9276-0/TABLES/1

Van Gulick, R. (2009). Functionalism. En A. Beckermann, B. P. McLaughlin, & S. Walter (Ed.), *The Oxford Handbook of Philosophy of Mind* (p. 128-151). Oxford University Press. https://doi.org/10.1093/OXFORDHB/9780199262618.003.0008

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2023). *Attention Is All You Need*.

Véliz, C. (2021). Moral zombies: why algorithms are not moral agents. *AI and Society*, *36*(2), 487-497. https://doi.org/10.1007/S00146-021-01189-X/METRICS

Vellino, A. (1985). Artificial intelligence: The very idea. *Artificial Intelligence*, *29*, 349–353.

Verbeek, P. P. (2008). Morality in Design: Design Ethics and the Morality of Technological Artifacts. En P. Kroes, P. E. Vermaas, A. Light, & S. A. Moore (Ed.), *Philosophy and Design* (p. 91-103). Springer. https://doi.org/10.1007/978-1-4020-6591-0_7

Verbeek, P.-Paul. (2011). *Moralizing technology : understanding and designing the morality of things*.

Vincent, J. (2019, novembre 27). *Former Go champion beaten by DeepMind retires after declaring AI invincible*. The Verge.

Vladimirov, N., Wang, C., Höckendorf, B., Pujala, A., Tanimoto, M., Mu, Y., Yang, C. T., Wittenbach, J. D., Freeman, J., Preibisch, S., Koyama, M., Keller, P. J., & Ahrens, M. B. (2018). Brain-wide circuit interrogation at the cellular level guided by online analysis of neuronal function. *Nature methods*, *15*(12), 1117-1125. https://doi.org/10.1038/S41592-018-0221-X

Wachter, S., Mittelstadt, B., & Russell, C. (2018). Counterfactual explanations without opening the black box: automated decisions and the GDPR. *Harvard Journal of Law and Technology*, *31*(2), 841-887.

Wallach, W., & Allen, C. (2009). Moral Machines: Teaching Robots Right from Wrong. *Moral Machines: Teaching Robots Right from Wrong*, 1-288. https://doi.org/10.1093/ACPROF:OSO/9780195374049.001.0001

Wittgenstein, L. (2012a). *Investigaciones filosóficas*. Crítica.

Wittgenstein, L. (2012b). *Tractatus logico-philosophicus - Alianza Editorial*. Alianza editorial .

Zachlod, D., Palomero-Gallagher, N., Dickscheid, T., & Amunts, K. (2023). Mapping Cytoarchitectonics and Receptor Architectonics to Understand Brain Function and

Connectivity. *Biological Psychiatry*, *93*(5), 471-479. https://doi.org/10.1016/J.BIOPSYCH.2022.09.014