UAB
Universitat Autònoma de Barcelona

Contra una etologia digital
Anàlisi dels usos artificials de la metàforacomputacional

Marc Oriol Crespí Jiménez

Contra una etologia digital

Anàlisi dels usos artificials de la metàfora computacional

Autor: Marc Oriol Crespí Jiménez
Director i Tutor acadèmic: David Casacuberta Sevilla
1308 - Programa de Doctorat en Filosofia
Departament de Filosofia
Universitat Autònoma de Barcelona
Any de dipòsit: 2024

Per a l'Àlex,

que hauria d'haver pogut llegir això

Abstract

Within the framework of the philosophy of digital technology, this work investigates how some of the leading figures in Artificial Intelligence, such as Stuart Russell, Nick Bostrom, David Chalmers, Gary Marcus, Melanie Mitchell, and Rodney Brooks, create a digital ethology. The analysis finds similarities between some of their statements and ethology, and studies their level of impact based on their degree of credibility. It analyses the conceptual requirements of digital ethology and relates them to a more classical figure, such as the computational metaphor. The conclusion states that this digital ethology becomes particularly dangerous when we invert the computational metaphor. This inversion puts humans in a situation of dependence on digital technology, prioritizing its proliferation as if it were the colonization of a new species.

Keywords: digital technology, digital ethology, artificial intelligence, computational metaphor, digital colonization, new species.

Resum

En el marc de la filosofia de la tecnologia digital, en aquest treball s'investiga la confecció d'una etologia digital per part d'algunes de les figures principals de la Intel·ligència Artificial, com Stuart Russell, Nick Bostrom, David Chalmers, Gary Marcus, Melanie Mitchell i Rodney Brooks. S'observen similituds d'algunes de les seves afirmacions amb una etologia i s'estudia el seu nivell d'impacte tenint en compte el seu grau de credibilitat. S'analitzen els requisits conceptuals de l'etologia digital i es relacionen amb una figura més clàssica, com és la metàfora computacional. Es conclou que aquesta etologia digital és especialment perillosa quan es produeix una inversió de la metàfora computacional. Aquesta inversió posa els humans en situació de dependència amb la tecnologia digital prioritzant la seva proliferació com si es tractés de la colonització d'una nova espècie.

Paraules clau: tecnologia digital, etologia digital, intel·ligència artificial, metàfora computacional, colonització digital, nova espècie.

Índex de continguts

«The price of metaphor is eternal vigilance»

A. Rosenblueth i N. Wiener

Presentació

Aquest treball és una etologia, l'etologia dels defensors de l'etologia digital. Per etologia digital, com s'explicarà detalladament en el primer capítol, s'entén un discurs que descriu i tracta el comportament d'entitats digitals com si fossin éssers vius. Per això, aquests discursos solen sorgir d'investigadors en el mal anomenat camp de la Intel·ligència Artificial (IA). La idea d'analitzar aquest fenomen des d'aquesta perspectiva sorgeix després de temps dedicat a la implementació de programes informàtics i a l'estudi teòric de la informàtica, en aquest ordre. Segurament, com Joseph Weizenbaum, el fet de primer haver bragat amb la manca de pulcritud del codi informàtic de diversos programadors així com haver descobert que a vegades un algoritme funciona sense saber ben bé per què ("afegim +1 al final?"), deu haver condicionat aquesta imatge del món de la informàtica com un gran trasto, la recursivitat del qual mai pot ser legítimament anomenada melodia. Quan més alts són els altaveus que afirmen la seva glòria, més incredulitat em generen. Suposo que tampoc ajuda haver treballat escrivint discursos comercials per aconseguir col·locar el producte en bondadosos i, després, queixosos clients.

El treball es divideix en dues parts. Una primera part de quatre capítols, tres dels quals constitueixen pròpiament el treball de camp al ser l'anàlisi de 182 proposicions extretes de textos de vuit autors vinculats de diferent manera al projecte etològic digital. Per tant, aquesta primera part és un mostrari, ni orientat ni exhaustiu, de diferents posicions agrupades en dos grans blocs: els autors de la por i els autors escèptics. I una segona part, de dos capítols, a la qual s'hi accedeix després d'un *intermezzo*, capítol breu que fa de frontissa lleugera entre el treball de camp, objectiu de la primera part, i l'anàlisi dels requisits conceptuals d'una etologia digital, objectiu de la segona part.

En el capítol 1 es presenten dues aproximacions diferents als textos que fonamenten una etologia digital. Per una banda, una d'intuïtiva, és a dir, immediata, instintiva, que no pretén basar-se en cap raonament; i, per l'altra, una de raonada, que s'assenta en els pilars clàssics de l'etologia, tal i com els va descriure Konrad Lorenz en la seva obra titulada *Fonaments de l'etologia* (1978). També es planteja una metodologia d'anàlisi d'aquestes textos a partir de dues variables: la perillositat i la credibilitat.

El capítol 2 està dedicat als autors de la por, és a dir, aquells que prioritzen un discurs catastrofista i distòpic davant dels nous programes que s'alliberen (una paraula amb connotacions etològiques i que es pot evitar fàcilment per *comerciar*) fent ús d'IA. S'han triat tres autors representatius d'aquest sector i, al mateix temps, amb posicions diferents i igualment consolidades: Stuart Russell, Nick Bostrom i David Chalmers.

El capítol 3 està dedicat a dues cartes, una de les quals va causar grans titulars l'any 2023 fins al punt de tapar la primera, que s'havia publicat un dia abans. La famosa és la que van signar Elon Musk i companyia defensant una aturada de 6 mesos de tots els projectes d'IA davant del greu perill que representaven. L'altra, un escrit de Bill Gates al seu blog l'optimisme del qual va quedar completament eclipsat. Són dos documents que comparteixen format i intencions, però no manera de fer, i per això resulta interessant comparar-los.

El capítol 4 està dedicat als autors escèptics, si més no escèptics amb el discurs de la por. No tenen una actitud estrictament optimista, sinó que reclamen certa rigorositat i seriositat a tots els actors del camp de la IA. Es tracta de Gary Marcus, Melanie Mitchell i Rodney Brooks, cada un d'ells amb els seus matisos, però tres figures crítiques amb tota la part pantomímica d'una etologia digital.

Després d'aquest mostrari ordenat, hi ha l'*intermezzo*, que serveix de frontissa per bolcar el treball cap a una part més conceptual, per això primer agrupa les conclusions de la primera part, després presenta l'assumpte de l'etologia digital des d'una perspectiva del llenguatge i, finalment, vincula els problemes de la filosofia de la tecnologia digital amb els problemes de la filosofia de tota la vida. Perquè la filosofia de la tecnologia digital no deixa de ser filosofia, però centrada en artefactes digitals.

En el capítol 5 s'analitza la metàfora computacional com un dels requisits conceptuals principals d'una etologia digital: se'n busquen els antecedents i se'n ressegueixen les diferents formulacions fins al computacionalisme, que pretén deixar-la enrere per abraçar la seva inversió, figura que s'ha anomenat la metàfora del programador (o de l'enginyer informàtic). Aquest recull històric comença amb investigadors de la lingüística, la filosofia i la ciència en general, com Max Black, Richard Boyd i Thomas Kuhn, que discuteixen sobre el paper de les metàfores en la ciència; després, s'especialitza en autors de l'àmbit de la informàtica que explícitament han tractat el tema de la metàfora computacional, com Norbert Wiener, Warren McCulloch, Claude Shannon, John von Neumann i Joseph Weizenbaum; finalment, acaba amb la interpretació del sentit i límits de la metàfora computacional de Daniel Dennett i de Steven Pinker. La vessant biològica de la metàfora

computacional, és a dir, la seva contribució a una etologia digital, s'analitza des de l'òptica de John Dupré.

El capítol 6 pretén ser una recopilació de resultats del treball de camp en conjunció amb els requisits conceptuals derivats de la metàfora computacional i la seva inversió. El treball es tanca amb un exemple d'etologia digital que ha afectat i segueix afectant les aules d'escoles i instituts i que té en el *multitasking* un dels seus emblemes flagrants. Aquest discurs és contraposat a les crítiques i alternatives presentades per Michel Desmurget, Roberto Casati i, novament, Joseph Weizenbaum.

En resum, l'objectiu d'aquests sis capítols és mostrar com es poden entendre des d'una perspectiva etològica una sèrie de discursos que s'estan donant, no només en un entorn comercial, sinó també en un d'acadèmic. I, en segon terme, connectar aquest fenomen amb la metàfora computacional i, per extensió, amb certs problemes clàssics de la filosofia.

Primera part

> By attributing all behavior under the sun to a single learning mechanism, behaviorism set up its own downfall. Its dogmatic overreach made it more like a religion than a scientific approach. Ethologists loved to slam it, saying that instead of domesticating white rats in order to make them suitable to a particular testing paradigm, behaviorists should have done the opposite. They should have invented paradigms that fit "real" animals.
>
> Frans de Waal, *Are we Smart Enough to Know how Smart Animals Are?*

1. Què és una etologia digital?

1.1 Una aproximació intuïtiva a mode d'introducció

Aquest treball identifica l'aparent oxímoron *etologia digital* amb tots aquells textos, discursos o proposicions que tracten els objectes digitals com si fossin un ésser viu. Al nostre voltant en tenim diferents exemples, des dels més innocents i fruit de l'habitual antropomorfització (quan diem que l'ordinador està pensant) fins a més ambigus (com quan assumim que certa IA ha pintat un quadre o ha guanyat una partida de go). El més explícit i recent, les declaracions de Blake Lemoine sobre LaMDA (una IA de Google) en les que defensa, després de diversos intercanvis de preguntes i respostes amb aquest *xatbot*, que és un ésser que sent: «LaMDA is a sweet kid who just wants to help the world be a better place for all of us. Please take care of it well in my absence».[1]

Hi ha diferents mecanismes que contribueixen en major o menor mesura a la creació d'una etologia digital: l'ús antropomòrfic de termes aplicant-los a ens digitals (i.e., l'ordinador està pensant); l'ús de metàfores o comparacions entre elements analògics i digitals (i.e., l'ordinador funciona com un cervell); descriure el comportament d'un ésser viu amb vocabulari propi d'un ens digital (i.e., els nens són *multitasking*); disminuir capacitats o atributs d'uns ésser viu per tal que un ens digital s'hi pugui comparar (i.e. els humans tampoc som tan intel·ligents com ens pensem); la desvinculació de l'eina respecte el seu creador, sense que això representi necessàriament una amenaça (i.e. l'aparició de la IA); apel·lar a la inexplicabilitat del seu funcionament (i.e., no sabem com ho fan); l'assumpció que els humans, de fet, són màquines (i.e., el cervell literalment és un computador); presentar la IA com la nova pedra filosofal que resoldrà els problemes de la humanitat (i.e., la IA resoldrà el canvi climàtic); o directament la presentació de la IA com una nova espècie

1 TIKU, Nitasha (11.06.2022). "The Google engineer who thinks the company's AI has come to life" en *The Washington Post*. Consultat el 13 de juny de 2022 a: https://www.washingtonpost.com/technology/2022/06/11/google-ai-lamda-blake-lemoine/

invasora amb la qual cal negociar i entendre-s'hi, ja que és una amenaça (i.e., els humans estem en risc d'extinció a causa de la IA).

No totes aquestes maneres sembla que impliquin exactament el mateix, ja sigui perquè alguns usos estan més incorporats a la llengua que altres, ja sigui pel context en què s'emet cada un d'aquests missatges i la seva intencionalitat. Tanmateix, des d'una perspectiva més aviat intuïtiva, sembla que algunes expressions pretenen quelcom més que descriure una situació. Així, sembla completament innocu afirmar que l'ordinador està pensant, mentre que sembla més interessat, fins el punt de preguntar-se què volen vendre, quan diuen que els humans estem en risc d'extinció a causa de la IA.

El projecte d'analitzar en què consisteix una etologia digital, exemplificar i criticar-ne els seus usos, així com intentar esbrinar els requeriments conceptuals que la permeten, caldrà que comenci mostrant fins a quin punt certs discursos actuals es basen en la teoria etològica. Per això, el millor és anar a l'inici de l'etologia.

1.2 Una aproximació raonada a mode de contextualització

Etimològicament, *etologia* és un terme compost per *ethos* (caràcter) i *logos* (estudi), i inicialment es va fer servir per referir-se a l'estudi del caràcter humà. En aquest sentit la fa servir, per exemple, John Stuart Mill el 1843 quan titula el capítol 5 del seu sisè llibre de *A System of Logic* "Of Ethology, or the Science of the Formation of Character". És ja Mill qui proposa diferenciar l'etologia de la psicologia de la següent forma:

> The name is perhaps etymologically applicable to the entire science of our mental and moral nature; but if, as is usual and convenient, we employ the name Psychology for the science of the elementary laws of mind, Ethology will serve for the ulterior science which determines the kind of character produced in conformity to those general laws, by any set of circumstances, physical and moral.[2]

Mentre que la voluntat de diferenciar-se de la psicologia va esdevenir una constant de l'etologia, el seu camp d'estudi no: la proposta de Mill no va tenir continuïtat i no va ser fins el 1902, quan William Morton Wheeler (un mirmecòleg americà) va utilitzar el terme per parlar del comportament social de les formigues, que va començar a significar el que canònicament s'entén avui en dia: «Ethology is the science that study animal behaviour, usually with a focus on behaviour under natural conditions»[3]. Precisament, aquesta idea que calia observar la conducta dels animals en el seu entorn és un dels trets diferencials que durant el segle XX més va enfrontar les propostes que

2 MILL, John Stuart (1843). *A System of Logic*. Consultat el 29 de juliol de 2023 a: https://www.laits.utexas.edu/poltheory/mill/sol/sol.b06.c05.html

venien de la psicologia comparada americana, més partidaris de preparar proves altament controlables en un laboratori, de les propostes europees que sorgien a l'entorn de la zoologia[4], com les de Konrad Lorenz, Nikolassa Tinbergen i Karl von Frish (mereixedors del premi Nobel de Fisiologia o Medicina de 1973).

Una de les constants també d'aquesta nova ciència ha estat aconseguir una metodologia que pogués ser considerada pròpiament científica, a diferència de la tradicional descripció del comportament animal que s'havia fet i que es pot llegir en textos clàssics com *Investigació sobre els animals* d'Aristòtil. Una altra constant ha estat la pugna i posterior intent de síntesi entre, per una banda, autors que apel·laven a l'instint animal, per tant, a trets innats o genètics (Timbergen titula una de les seves obres principals *The Study of Instinct*), i autors que apel·laven a l'aprenentatge com a base del comportament per l'altra (un exemple clàssic és l'obra de J.B. Watson, *Behaviourism*, autor influenciat pel condicionament clàssic de Pavlov[5]). Aquestes diferències prenen altres formes, derivades directament d'aquesta disputa, al defensar aquests segons una proposta mecanicista (de l'estil del comportament condicionat per recompensa de Skinner) mentre que els primers advocaran majoritàriament per una proposta vitalista. Arran d'aquestes diferències, també ha estat constant la preocupació per aconseguir descriure el comportament animal sense que les seves proposicions prenguessin, encara que fos implícitament, un sentit teleològic i, al mateix temps, sense renunciar a cert ús antropomòrfic inherent al tractament dels humans com un animal més. El problema rau en intentar trobar un equilibri entre un discurs com el de la sociologia (que té en compte els propòsits humans), amb un discurs com el de la biologia (que assumeix que l'evolució no té cap propòsit en si mateixa), especialment quan l'estudi consisteix en la comparació del comportament entre diferents espècies. Konrad Lorenz (1903-1989) ho contextualitza amb un exemple de com cal entendre la pregunta del "per a què" en aquest camp:

> Cuando preguntamos: "¿para qué el gato tiene uñas puntiagudas curvas y retráctiles?", y contestamos brevemente: "para cazar ratones", esta pregunta no significa en manera alguna que creamos en una predeterminación inherente al universo y a la evolución orgánica. En verdad es

3 FERICEAN , Mihaela Liana; RADA , Olga; BADILITA, Mihaela (2015). "The history and development of ethology" en *Research Journal of Agricultural Science,* 47 (2), 2015, pàg. 45.

4 PELÁEZ del Hierro, Fernando; VEÀ BARÓ, Joaquim (1997). *Etología. Bases biológicas de la conducta animal y humana*, Madrid, Ediciones Pirámide, 1997, pàg. 19.

5 FERICEAN , Mihaela Liana; RADA , Olga; BADILITA, Mihaela (2015). "The history and development of ethology" en *Research Journal of Agricultural Science,* 47 (2), 2015, pàg. 46.

una abreviación de la pregunta: "qué función específica es aquella cuyo valor de preservación de la especie confirió a los animales carnívoros (*Felidae*) esa peculiar forma de sus garras?".[6]

En aquesta mateixa obra, Lorenz recorda la diferència introduïda pel botànic Colin Pittendrigh (1918-1996) entre teleologia i teleonomia, terme que designaria l'aparent finalitat de l'adaptació per a la conservació de les espècies, però sense l'aura mística que connota una teleologia. Contra aquesta proposta, Ernst Mayr (1904-2005), biòleg evolutiu, va defensar que es restringís l'ús del terme *teleonomia* als sistemes que operen en base a un programa d'informació codificada, és a dir, al camp de la tecnologia, en el qual el propòsit de cada invent està pensat pel seu inventor. Tanmateix, la polèmica de fins a quin punt es pot fer un discurs finalista quan es descriu el comportament d'una espècie no pot plantejar cap problema dins l'estudi de la tecnologia, on és obvi que un programa només funciona bé quan fa allò que volia que fes el seu programador. Per tant, mentre que és lícit i, fins i tot, òptim fer un discurs teleològic o finalista dins del camp de la invenció tecnològica, com la programació i la IA, no ho és dins dels camps de les ciències naturals, ja que s'estaria suposant un propòsit en l'evolució. El que no és lícit, ni òptim, en el camp de la tecnologia és traslladar aquest discurs teleològic cap al camp del qual s'han agafat una sèrie de conceptes prestats com és el camp de la biologia (un dels conceptes prestats és *evolució*). Per tant, resultaria tan contradictori intentar fer un discurs teleològic en biologia (sense defensar un disseny intel·ligent) com fer un discurs evolutiu en tecnologia (sense defensar una divinització humana). Tanmateix, això mateix és el que proposa, per exemple, David Chalmers en aquesta conversa: «There's no contradiction between adaptation and computation»[7]. Aquest tipus de contradiccions és a les que s'arriba quan es pretén fer una etologia digital, oxímoron en si mateix.

Per mostrar com certs discursos tecnològics cauen en una etologia dels objectes digitals, es mostrarà com algunes estratègies pròpies de l'etologia s'apliquen al camp de la tecnologia sense un motiu ni suficientment justificat ni raonable ni exempte de perills. Per fer-ho, se seguirà, pel seu aspecte clàssic, la manera d'investigar de l'etologia tal i com la descriu Lorenz en el capítol II de *Fonaments de l'etologia*, titulat "El mode de recerca de la biologia, en particular, el de l'etologia", qui defensa que l'etologia té deu trets distintius que la diferencien, no només de la física, sinó també de la biologia. Aquests trets són els següents (el títol de cada una d'elles coincideix amb el títol que assigna Lorenz i quan cal, entre parèntesi, un aclariment):

6 LORENZ, Konrad (1978). *Fundamentos de la etología. Estudio comparado de las conductas*, Barcelona, Ediciones Paidós, 1986, pàgs. 41-42. La cursiva, així com la resta de cursives dins una citació, són en l'original.

7 BROOKS, Rodney A., e*t alii* (13.05.2019). "The Cul-de-Sac of the Computational Metaphor A Talk By Rodney A. Brooks" en *Edge*. Consultat el 30 de juliol de 2023 a: https://www.edge.org/conversation/rodney_a_brooks-the-cul-de-sac-of-the-computational-metaphor

1. **El concepte de sistema o totalitat:** «El objetivo del biólogo es volver comprensible un sistema orgánico como un *todo*»[8]. En la investigació etològica, la totalitat ja ve donada pel món natural: l'ésser viu que s'estudia és sencer, és un animal o una planta, i el biòleg busca comprendre les relacions causals de i entre totes les seves parts dins d'aquest tot harmònic. En canvi, en el món de la tecnologia, no hi ha realitat donada, ja que precisament el que es vol és fer un invent, és a dir, obtenir una cosa que per definició no existeix. Tanmateix, es fa certa extrapolació d'aquesta idea al món de la tecnologia digital quan enlloc de pretendre resoldre un problema concret (i.e., subjectar un objecte a la paret) inventant una eina concreta (i.e., un tornavís i un vis), el programador ha d'introduir un propòsit global a la seva creació, ha de demostrar que el codi que ha fet és escalable en un tot imaginari (i.e., un algoritme d'autoaprenentatge ha de poder aprendre de tot, no només a reconèixer un tipus d'imatges, sinó totes les imatges possibles; no només jugar a un sol joc, sinó a tots els jocs possibles, encara que, estrictament parlant, ni estigui reconeixent imatges (simplement les classifica a partir d'una coincidència de bits), ni estigui jugant a cap joc en concret (sinó aplicant combinacions de patrons a partir del càlcul de probabilitat d'encert). Tanmateix, tal i com reconeix Stuart Russell, un dels investigadors més reputats en aquest camp, la investigació de la IA és total o no és: «That's the ultimate goal of AI research: a system that needs no problem-specific engineering and can simply be asked to teach a molecular biology class or run a government».[9]

2. **La successió dels passos del coneixement dictada pel caràcter dels sistemes (o l'ordre de successió el dicta el tot):** si l'objectiu del biòleg és comprendre el tot, és a dir, el sistema com a unitat, cal primer observar aquesta totalitat com a donadora de sentit final (*top-down*), ja que en cas contrari (*bottom-up*) les parts podrien redirigir el sentit del tot. Lorenz ho diu en negatiu: «Por consiguiente cometemos en principio un error metodológico si aislamos experimental o incluso idealmente una conexión causal y la investigamos en un solo sentido»[10]. Aquesta consigna no sembla tenir cap sentit en el món tecnològic digital on la invenció pot sorgir de la mera experimentació amb l'eficiència d'un algoritme o de l'encert casual dels seus *outputs*, com en el mencionat etiquetatge d'imatges, la identificació de les quals no es basa en característiques perceptibles pels humans, com explica Melanie

8 *Ibídem*, pàg. 48.

9 RUSSELL, Stuart (2019). *Human Compatible. AI and the Problem of Control*, Londres, Penguin Books UK, 2020, pàg. 46.

10 LORENZ, Konrad (1978). *Fundamentos de la etología. Estudio comparado de las conductas*, Barcelona, Ediciones Paidós, 1986, pàg. 49.

Mitchell, una autora crítica amb alguns d'aquests discursos etològics digitals: «This is because such systems are susceptible to *shortcuts learning*: learning statistical associations in the training data that allow the machine to produce correct answers but sometimes for the wrong reasons»[11]. És a dir, tot i que qualsevol algoritme té un propòsit fixat pel seu programador, la troballa d'una millora d'una de les seves parts pot alterar aquest propòsit fixat i convertir un algoritme inicialment pensat per millorar la velocitat d'un videojoc en una peça clau dels sistemes d'aprenentatge profund, com reconeix Gary Marcus: «GPUs were developed for video games and other graphics applications, not for training deep learning systems or mining crypto-currency»[12]. Per tant, no hauria de suposar cap problema dins del discurs tecnològic admetre que una peça pot canviar el funcionament del tot, en la mesura que en tecnologia hauria de prevaldre l'optimització funcional per sobre de la factualitat de la realitat. I, de fet, sol passar així, com recorda James Lanier quan parla del fitxers: «The first iteration of the Macintosh, which never shipped, didn't have files. […] UNIX had files; the Mac as it shipped had files; Windows had file. Files are now part of life»[13]. Ara bé, aquesta realitat de la tecnologia (que és una diferència clau de la biologia) no encaixa gens bé amb el discurs etològic digital: per tal d'imposar la creença que l'ens digital és un ésser viu, negaran la possibilitat que una part hagi pogut modificar el tot. Irònicament, això contradiu algun dels postulats de l'evolució, com reconeix el mateix Lorenz:

> Por tanto, nunca la estructura total de un organismo es igual a una obra humana diseñada por un constructor largamente previsor según un único plano, sino que se parece mucho más a la casa que un colono ha construido para sí: primero levanta una simple cabaña para protegerse contra el viento y la lluvia, agrandándola a media que aumentan posesiones y familia. La cabaña inicial no es derruida, sino que se convierte en trasero, y casa todos los cuartos de la construcción total pierden su función original en el curso de la evolución. Los restos históricos reconocibles como tales se conservan por el mismo hecho que determinó que la construcción nunca pudiera ser completamente derribada ni planificada de nuevo: ello era

11 MITCHELL, Melanie (26.04.2021). "Why AI is Harder Than We Think", en *arXiv*. Consultat el 30 de juliol de 2023 a: https://arxiv.org/abs/2104.12871

12 MARCUS, Gary; DAVIS, Ernst (18.10.2023). "Reports of the birth of AGI are greatly exaggerated" en *Marcus on AI.* Consultat el 5 de juliol de 2024 a: https://garymarcus.substack.com/p/reports-of-the-birth-of-agi-are-greatly?publication_id=888615&post_id=138077794&isFreemail=true&r=2e3aia

13 LANIER, Jaron (2010). *You are not a gadget. A manifesto*, Londres, Penguin Books, 2011, pàg. 12.

> imposible precisamente porque estuvo habitada todo el tiempo y sometida a un uso intenso.[14]

Segurament, d'aquesta diferència entre una construcció humana i un organisme natural, i de voler ignorar-la, neixen la major part de les contradiccions d'una etologia digital.

3. **La capacitat cognitiva de la percepció (o la mirada preval per sobre de la visió):** en un camp en què l'observació dels animals en el seu hàbitat és imprescindible, cal trobar mecanismes per a què aquesta no perdi valor o sigui qüestionada per paràmetres excessivament reduccionistes propis d'altres camps com la química, en el qual els experiments poden i han de ser reproduïbles en condicions normals. Aquí la mirada de l'investigador de camp s'ha de guanyar la validesa per sobre del mesurament quantificat del laboratori. Lorenz pretén justificar aquesta validesa fent servir una estratègia comparativa: «Ni la abeja ni el hombre están interesados en establecer las longitudes de onda absolutas de la luz. Lo que deben poder hacer es reconocer un objeto biológicamente relevante por sus propiedades de reflexión»[15]. En el seu intent de posar en valor aquesta mirada contraposa la mera racionalitat que pot aportar una computadora a la capacitat de descobrir lleis imprevistes de la mirada global: «La investigación racional y, tanto más, las computadoras sólo pueden contestar preguntas que ya hayan sido claramente planteadas por la razón humana»[16]. Per combatre aquesta idea i poder fer una etologia digital caldrà poder presentar la tecnologia com a creativa, descobridora de noves lleis, obviant la feina de configuració prèvia de personal especialitzat en la branca que la IA hagi de treballar, tal com explica detalladament Kate Crawford en *Atlas of AI*[17]. Una forma de fer-ho consistirà en amagar que la base dels algoritmes d'aprenentatge automàtic és una funció estocàstica (el mètode de Montecarlo o derivats), és a dir, que utilitza fórmules basades en l'atzar per optimitzar la cerca de patrons, cosa que és l'antítesi de la mirada humana, que els que busca són expressions de sentit. Per tant, una IA podrà canviar la visió quantificadora per una mirada quan aquesta sigui una mímica de la mirada, això és, el resultat estadístic de la mirada.

14 LORENZ, Konrad (1978). *Fundamentos de la etología. Estudio comparado de las conductas*, Barcelona, Ediciones Paidós, 1986, pàg. 41.

15 LORENZ, Konrad (1978). *Fundamentos de la etología. Estudio comparado de las conductas*, Barcelona, Ediciones Paidós, 1986, pàgs. 52-53.

16 *Ibídem*, pàg. 55.

17 CRAWFORD, Kate (2021). *Atlas of AI. Power, Politic, and the Planetary Costs of Artificial Intelligence*, New Haven, Yale University Press, 2021.

Explicar això de forma clara, explicitar, per exemple, que l'encert de la construcció d'una frase com a *output* a una pregunta en un ChatGPT es basa en un resultat estadístic, comprometria la possibilitat d'escriure una etologia digital –de fet, li va costar la feina a Timnit Gebru quan va firmar, juntament amb Emily M. Bender, l'article titulat "On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?"[18]. Tot i que això pugui semblar anecdòtic, una de les qüestions claus perquè es pugui afirmar que hi ha intel·ligència, és si hi ha comprensió[19]. Amagar que això no passa en un LLM (Large Language Models, base de la IA actual) és essencial per poder estendre la idea que hi ha realment una *intel·ligència* artificial.

4. **La denominada afició (o l'afició és una virtut)**: com en qualsevol camp, també en l'etologia és valorable tenir certa afició pel que s'investiga, altrament, les hores necessàries per observar el comportament animal poden ser molt dures. És obvi que aquesta virtut també es dona en la programació, amb l'afegit que aquesta té un component addictiu, com va assenyalar Jospeh Weizenbaum, qui el 1966 fou el primer programador d'un llenguatge, ELIZA, que responia en anglès a preguntes fetes en anglès. Tot i que no abunden els estudis específics entre la programació i l'addicció, hi ha dades que demostren que les pantalles, en general, produeixen addicció, com ha defensat Michel Desmurget: «También es difícil para los propios menores reconocerse en semejante arquetipo [el de l'addicte] especialmente porque con la dependencia a los dispositivos digitales ocurre lo mismo que con las demás adicciones: la negación es tenaz y frecuente»[20]. Per tant, mentre que l'afició d'observar animals no produeix cap addicció per ella mateixa, l'afició d'observar entitats digitals, si més no en una pantalla, sí que la produeix, fins i tot quan no hi ha una voluntat explícita per part de les persones que l'han programada. Així, constituirà una contribució a una etologia digital el fet de negar aquesta addicció i, per contra, presentar-la com una virtut.

5. **L'observació d'animals en llibertat i en captivitat:** tot i que cal observar els animals en llibertat abans d'emetre cap judici, aquest també pot contrastar-se més acuradament si posteriorment es fan proves en captivitat, molt especialment pel que fa a l'estudi de

18 BENDER, Emily M.; GEBRU, Timnit (01.03.2021). "On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?" en *FAccT '21: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, març 2021, pàgs. 610–623. Consultat el 30 de juliol de 2023 a: https://doi.org/10.1145/3442188.3445922.

19 Un dels principals camps d'investigació en IA és el reconeixement i comprensió de textos i imatges.

20 DESMURGET, Michel (2019). *La fábrica de cretinos digitales*, Barcelona, Planeta, 2023, pàg. 207.

patologies. Ara bé, el procés no es pot invertir, ja que la conducta en captivitat pot estar alterada per la mateixa captivitat, cosa que restaria valor a qualsevol observació. A nivell tecnològic, aquesta mateixa afirmació és la que es segueix quan s'*allibera* (l'ús d'aquest terme també és significatiu) un nou programa, moment indispensable si es vol comprovar que tot el que s'ha observat en laboratori realment ocorre en el seu espai natural, és a dir, el mercat. De fet, han passat a la història l'alliberament d'algunes aplicacions que han hagut de retirar-se del mercat a les 72 hores davant el perill que generava la seva actuació, com, per exemple, va fer Meta amb Galactica, una IA entrenada exclusivament amb articles científics que va començar a fer afirmacions poc contrastades i un gavadal de sense sentit[21]. També s'ha descrit i tipificat com a *model drift* (que es podria traduir com una degradació o desviació del model) el fenomen que sol ocórrer al cap d'uns mesos d'haver posat en servei un LLM: aquest comença a perdre els nivells de fiabilitat, fossin quins fossin, inicials[22]. Per tant, és imprescindible, si es vol construir una veritable etologia digital, que un programa desenvolupat necessàriament en un laboratori sigui alliberat i pugui ser observat en llibertat, ja que en cas contrari, aparentment, el programa no adquireix la seva màxima expressió. I constituiria el súmmum etològic si, després d'una observació exhaustiva, es capturés de nou el programa i se li fessin proves per tastar si el seu comportament havia estat programat o adquirit durant la seva vida lliure, cosa que permetria estudiar-ne en el laboratori els possibles comportaments patològics per tal de refinar-los per a una segona versió. De fet, això és el que ha fet, per exemple, OpenAI amb el ChatGPT: obrir-lo per tal que pugui ser entrenat entre el màxim de població i així poder-ne depurar els errors, tal i com demanen els seus creadors en la nota de llançament: «There's still a lot of work to do, and we look forward to improving this model through the collective efforts of the community building on top of, exploring, and contributing to the model».[23]

21 HEAVEN, Will Douglas (18.11.2022). "Why Meta's latest large language model survived only three days online" en *MIT Technological Review*. Consultat el 13 d'agost de 2023 a: https://www.technologyreview.com/2022/11/18/1063487/meta-large-language-model-ai-only-survived-three-days-gpt-3-science/

22 NELSON, K.; CORBIN, G.; ANANIA, M.; KOVACS, M.; TOBIAS, J.; and BLOWERS, M. (2015). "Evaluating model drift in machine learning algorithms" en *IEEE Symposium on Computational Intelligence for Security and Defense Applications (CISDA)*, Verona, NY, USA, 2015, pàgs. 1-8. Consultat l'11 de juliol de 2024 a: https://ieeexplore.ieee.org/document/7208643

23 OpenAI (14.03.2023). "GPT-4" en *OpenAi*. Consultat el 13 d'agost de 2023 a: https://openai.com/research/gpt-4.

6. **L'observació d'animals domesticats en llibertat:** pels etòlegs, alliberar animals domesticats pot ser útil per comprovar si un comportament és innat o adquirit, especialment quan aquest comportament està relacionat amb la sociabilitat. En l'àmbit tecnològic, tota entitat és, en aquest sentit, domesticada, però com s'ha vist en l'apartat anterior, deixar-la en llibertat també pot aportar informació interessant, per exemple, per esbrinar la causa d'un biaix. La discussió sobre si els biaixos són adquirits (entrenament amb dades esbiaixades d'origen) o innats (biaix en el mateix codi d'aprenentatge) no està resolt, doncs tot i que sembli evident que depenent del conjunt de dades amb les quals s'entreni un algoritme amb IA pot modificar el seu comportament i mostrar un tipus de biaix social (*societal bias)*, també és obvi que la preparació dels conjunts de dades (*design datasets*) i la seva modalització (*controlled datasets*) és responsabilitat dels enginyers que treballen també per corregir aquell comportament i evitar així un biaix estadístic (*statistical bias*)[24], tal i com va fer evident la disputa a través de Twitter entre Yann Lecunn i Tinnit Gebru[25]. De fet, trobar un equilibri en la modelització de les dades corregint biaixos a base d'incrementar la diversitat d'aquestes dades és complicat: «But it is not like more data diversity is always better; there is a tension here. When the neural network gets better at recognizing new things it hasn't seen, then it will become harder for it to recognize things it has already seen»[26]. Per tant, la clàssica discussió entre innat i adquirit també es reflectirà en els discursos de l'etologia digital, cosa que permetrà veure paral·lelismes entre l'estudi etològic i el tecnològic. Tanmateix, caldrà comprovar si aquestes característiques en el discurs tecnològic són heretades directament de l'etològic o, simplement, són inherents a una problemàtica global, com sembla suggerir l'antiguitat del dilema.

7. **El coneixement dels animals com a mètode d'investigació (o per què cal conèixer prèviament els espècimens):** «Para saber si una pauta conductual observada corresponde a la de una especie en su hábitat natural o si es patológica, el investigador tiene que conocer muy bien la especie estudiada»[27]. Aquest precepte, que sembla molt raonable quan es tracta de l'estudi d'animals, és complicat d'aplicar en l'estudi d'entitats digitals: no hi ha

24 KRISHNAMURTHY, Prabhakar (12.09.2019). "Understanding Data Bias" en *Towards Data Science*. Consultat el 15 d'agost de 2023 a: https://towardsdatascience.com/survey-d4f168791e57

25 SYNCED (30.06.2020). "Yann LeCun Quits Twitter Amid Acrimonious Exchanges on AI Bias" en *Synced*. Consultat el 15 d'agost de 2023 a: https://syncedreview.com/2020/06/30/yann-lecun-quits-twitter-amid-acrimonious-exchanges-on-ai-bias/

26 BOIX, Xavier; ZEWE, Adam (21.02.2022). "Can machine-learning models overcome biased datasets?" en *MIT News*. Consultat el 15 d'agost de 2023 a: https://news.mit.edu/2022/machine-learning-biased-data-0221

coneixement previ d'un algoritme a no ser que es reutilitzi el codi d'un altre algoritme, cosa que sol ser freqüent a nivell pràctic quan els programes superen certa longitud de codi i hi ha diversos programadors treballant-hi. Tanmateix, aquesta problemàtica pràctica i realista no és objecte d'estudi pel que fa a la qüestió etològica, ja que aquí es pot considerar que les respostes d'un algoritme amb IA que no són desitjables, són patològiques, independentment de si aquestes respostes foren observades amb anterioritat. Per tant, el concepte de patològic també s'haurà d'analitzar, ja que o bé és un concepte sense sentit si s'entén trivialment (tot comportament no desitjat és patològic) o bé és un tret etològic quan s'aplica a la tecnologia, ja que no pot existir una patologia en programació sinó una mala codificació. En aquest sentit, tractar les respostes esbiaixades d'un algoritme amb IA com a patològiques podrà ser considerat un tret etològic, al igual que amagar-les sota conceptes com *caixa negra*.

8. **L'experiment que sorgeix del tot orgànic (o per què l'experiment ha de tenir en compte l'ecologia):** Lorenz, novament, critica l'escissió sovint ideològica entre biòlegs de camp i biòlegs de laboratori (de fet, ell en diu biòlegs i psicòlegs), ja que allò ideal és contrastar una observació feta al camp amb l'estudi sistemàtic del laboratori, tal com expressa a través de les paraules citades de Fritz Knoll: «Las experiencias de laboratorio apropiadas y realizadas con toda la prudencia crítica necesaria suelen permitirnos averiguar los últimos matices de un fenómeno particular que se han sustraído a una clara comprensión en el demasiado articulado ambiente natural»[28]. Ara bé, a nivell tecnològic resulta confús com es pot dissenyar allò que Lorenz anomena una *das nicht-disruptive Experiment*, en la mesura que en aquest camp l'experiment sempre va abans de l'observació en el seu medi natural, és a dir, el mercat. Per tant, com en el punt anterior, només es podrà considerar un ús etològic si algú pretén que ha capturat un algoritme amb IA que campava lliurament i que, per contrastar quelcom observat, se li fan una sèrie de proves específicament dissenyades per testejar aquell comportament.
9. **L'experiment amb privació d'experiència:** Lorenz es planteja què es pot extreure d'un experiment explícitament dissenyat per privar d'una experiència concreta a un animal, per exemple, quan es priva de terra a un cadell i, tot i així, intenta enterrar un os i tapar-lo amb una terra imaginada. La resposta és que si, malgrat això, l'animal mostra certa conducta que es pot assegurar que no ha estat en cap cas adquirida, cal atribuir a aquesta conducta una

27 LORENZ, Konrad (1978). *Fundamentos de la etología. Estudio comparado de las conductas*, Barcelona, Ediciones Paidós, 1986, pàg. 63.

28 *Ibídem*, pàg. 66.

base filogenètica. Tanmateix, existeixen altres propostes per interpretar aquest fet, com la d'un altre pioner de l'etologia i creador del concepte de *Umwelt*, Jakob von Uexküll, que atribueix l'aparent sincronia entre la resposta innata d'un organisme i un entorn ecològic concret a una harmonia preestablerta, enlloc d'un mecanisme evolutiu d'adaptació. Aquesta harmonia preestablerta podria associar-se fàcilment a la idea de disseny intel·ligent que, en alguns investigadors en IA, sembla aparèixer quan fan analogies entre el disseny d'estratègies de programació i l'evolució. Per tant, també es considerarà una pràctica etològica l'equiparació entre disseny informàtic i naturalesa, en la mesura que aquesta només es pot fer a través de suposar un disseny intel·ligent de la naturalesa i, per tant, equipara el programador amb un dissenyador de l'univers (*law maker*), tendència que ja va denunciar també Joseph Weizenbaum el 1976:

> The computer programmer, however, is a creator of universes for which he alone is the lawgiver. So, of course, is the designer of any game. But universes of virtually unlimited complexity can be created in the form of computer programs. Moreover, and this is a crucial point, systems so formulated and elaborated act out their programmed scripts. They compliantly obey their laws and vividly exhibit their obedient behavior. No playwright, no stage director, no emperor, however powerful, has ever exercised such absolute authority to arrange a stage or a field of battle and to command such unswervingly dutiful actors or troops.[29]

10. **La part en relativa independència del tot (o de si es pot, millor començar per un punt fixe i independent del tot):** «Todos los manuales de anatomía empiezan con una descripción del esqueleto, y también en el estudio de la conducta fue legítimo iniciar los análisis por determinadas capacidades *no modificables* de sistema nervioso central»[30]. Aquest plantejament, dins del camp de la tecnologia, sembla que ha arrelat fortament en els autors més partidaris de seguir els passos de l'evolució, com el primer Rodney Brooks (ajudat per Daniel Dennett) i Hubert Dreyfus: fins que no es tingui clarament definit el funcionament de les parts més bàsiques i estables com són els comportament sensomotrius, no té massa sentit plantejar-se el funcionament d'altres parts vinculades a la intel·ligència: «on the basis of Brooks' success with insect-like devices, instead of trying to make, say, an

29 WEIZENBAUM, Joseph (1976). *Computer Power and Human Reason*, Londres, Penguin Books Ltd, New Ed, 1984, pàg. 115.

30 LORENZ, Konrad (1978). *Fundamentos de la etología. Estudio comparado de las conductas*, Barcelona, Ediciones Paidós, 1986, pàg. 74.

artificial spider, Brooks and Dennett decided to leap ahead and build a humanoid robot»[31]. Per tant, també es considerarà una proposta etològica aquells discursos que alineïn el disseny computacional amb l'evolució.

Resumint, es considerarà una col·laboració amb una etologia digital qualsevol dels següents comportaments dins de l'àmbit tecnològic i, concretament, digital (se segueix l'enumeració anterior):

1. La pretensió innecessària de totalitat.
2. La pretensió que aquesta totalitat té un sentit o direcció predefinida.
3. La pretensió que aquest sentit és natural.
4. La pretensió que aquesta naturalitat converteix en una afició innòcua el tracte amb les entitats digitals.
5. La pretensió que un programa ha de ser alliberat per gaudir de la seva màxima expressió.
6. La pretensió que l'actuació d'un programa alliberat permet extreure conclusions que sobrepassen al propi programa, com la de si hi ha trets adquirits o innats.
7. La pretensió que un programa informàtic pot tenir comportaments patològics i inesperats.
8. La pretensió que els comportaments patològics d'un programa agafen per sorpresa al seu programador.
9. La pretensió que el programador fa una feina similar a una deïtat, ja que crea entitats.
10. La pretensió que el programador ha de seguir els passos de l'evolució per aconseguir que el seu programa pugui inscriure's coherentment a la naturalesa.

L'anàlisi de l'ús d'aquests comportaments caldrà fer-lo en diferents tipologies de discursos, tant divulgatius com acadèmics, si es vol estudiar el fenomen de l'etologia digital a fons. De fet, aquest treball és en ell mateix una etologia dels divulgadors de la digitalització, cosa que la converteix en una veritable etologia digital, ja que, malgrat una etologia dels objectes digitals manqui de sentit, sí que pot resultar interessant una etologia de les persones que intenten construir-la així com les conseqüències que pot arribar a tenir segons el seu nivell etològic i impacte social o perillositat.

31 DREYFUS, Hubert L. (2007). "Why Heideggerian AI Failed and how Fixing it would Require making it more Heideggerian" en *Artificial Intelligence*, 171, 2007, pàg. 1141. Consultat el 6 de juliol de 2024 a: https://www.sciencedirect.com/science/article/pii/S0004370207001452

1.3 Els diferents nivells etològics i el seu impacte social

Al llarg de la història, diferents autors han estudiat l'impacte social de la tecnologia. Una de les obres clàssiques és la de *Technics & Civilization* (1934) de Lewis Mumford, en la qual, amb l'exemple del pas del repic de les campanes al rellotge exemplifica com ha influït, des dels seus inicis, la tecnologia del control i senyal del temps en els humans[32]. També Jacques Ellul a *La technique ou l'enjeu du siecle* (1954) denuncia com, depenent del període històric, aquesta influència pot ser més o menys beneficiosa: en ple capitalisme, qualsevol invent només és l'excusa pel següent, a través d'una dinàmica de problema–solució tècnica–nou problema (un dels exemples clàssics és el del desguàs desbrossador, que permet llençar les escombraries a la pica de la cuina a costa de pol·luir els rius, cosa que provoca la necessitat d'inventar depuradores d'aigua que, al seu temps, consumeixen grans recursos d'energia...[33]). Més recentment i recuperant una perspectiva heideggeriana que després influirà en algunes altres propostes (en concret la Dreyfus i la de Brooks), hi ha l'obra de Don Ihde titulada *Technology and the lifeworld* (1990). Ihde defensa que la tecnologia no és un fenomen independent dels humans, encara que tampoc en sigui exclusiu (alguns primats fan servir instruments). Així, la seva influència és recíproca: els humans, al crear-la, la influïm (en el sentit de donar-li forma des de dins); mentre que el seu ús acaba influint de tornada al comportament humà. Ho il·lustra amb l'adaptació d'una faula:

> Virtually every example of a human-technology interchange can illustrate this interrelation. Take the following gloss upon the fox and grapes story: The fox, seeing grapes too high to reach by his bodily jumping capacity, concludes that the grapes were sour; but the human, at first also unable to reach or jump to the grapes, picks up a stick and knocks the grapes down, thus not finding it necessary to conclude that the grapes are sour. Both fox and human, in the most narrow microperceptual sense, perceive the grapes as edible and desirable, but the primitive technological context made possible by the stick changes the perceptual sense of grapes as attainable and, with it, the macroperception the human may have both of the object of perception and of his or her ability to attain that object.[34]

Un dels deixebles conceptuals de Ihde és Peter-Paul Verbeek, juntament amb qui va desenvolupar el concepte d'aproximació post-fenomenològica dins la tradició de la filosofia de la tecnologia. Així ho defineix aquest darrer:

> All of these postphenomenological studies have at least two things in common. First of all, they study technology in terms of the relations between human beings and technological artifacts,

32 MUMFORD, Lewis (1934). *Técnica y civilización*, Logroño, Pepitas de calabaza Ed., 2020, pàgs. 37-44.

33 ELLUL, Jacques (1954). *La edad de la técnica*, Barcelona, Ediciones Octaedro, 2003, pàg. 98.

34 IHDE, Don (1990). *Technology and the lifeworld*, Indianapolis, Indiana University Press, 1990, pàg. 30.

> focusing on the various ways in which technologies help to shape relations between human beings and the world. They do not approach technologies as merely functional and instrumental objects, but as mediators of human experiences and practices. And second, they combine philosophical analysis with empirical investigation. Rather than 'applying' philosophical theories to technologies, the post-phenomenological approach takes actual technologies and technological developments as a starting point for philosophical analysis. Its philosophy of technology is in a sense a philosophy 'from' technology.[35]

En un dels articles publicats per Verbeek, juntament amb Nynke Tromp i Paul Hekkert i titulat "Deign for Socially Responsible Behavior: A classification of Influence Based on Intended User Experience", classificaven la tecnologia segons el canvi que pot produir en el comportament dels usuaris. Utilitzaven dues variables (força i saliència) que gradaven en els seus dos extrems respectius (força: *strong/weak*; saliència: apparent/hidden): «On the basis of two dimensions (i.e., salience and force), we classify four different types of influence: coercive, persuasive, seductive, and decisive influence»[36]. Inspirats per aquesta taula, aquí s'emula un eix de coordenades semblant que classifica els textos d'alguns autors segons la seva contribució a la confecció d'una etologia digital a partir de dues variables: la perillositat i la credibilitat. La classificació pretén analitzar la relació entre el nivell de perillositat o impacte social d'un conjunt de proposicions i el grau de credibilitat o honestedat o ingenuïtat o coherència (o manca d'aquests) del seu emissor, i com impacten aquests dos elements en la confecció d'una etologia digital. Es parteix de la idea que no totes les afirmacions col·laboren de la mateixa manera en la confecció d'una etologia digital, ja que depenent de quina informació es presenta, com es presenta i qui la presenta, es poden donar diversos nivells etològics, des dels més innocents i inofensius, als més mal intencionats i perillosos. Per tant, pel que fa al grau de perillositat o impacte social, es classifica en una escala que va del nivell inofensiu al perillós, i pel que fa al grau d'honestedat, en una escala que va de l'innocent o genuí al mal intencionat.

Es considera innocent quan l'emissor de les proposicions fa o bé un ús tècnic, clar i precís d'un terme o bé un ús clarament metafòric sense un aparent objectiu de confondre el receptor, ja sigui perquè no té cap interès evident en confondre'l (no hi guanya res d'aquesta confusió), ja sigui perquè, tot i tenir interessos evidents en confondre'l, fa tot el que pot per evitar-ho (per això

35 VERBEEK, Peter-Paul. "Postphenomenology" en *Peter-Paul Verbeek*. Consultat el 29 de juliol de 2024 a: https://ppverbeek.org/postphenomenology/

36 TROMP, Nynke; HEKKERT, Paul; VERBEEK, Peter-Paul (2011). "Deign for Socially Responsible Behavior: A classification of Influence Based on Intended User Experience" en *Design Issues.* Volume 27, Number 3, Summer 2011, pàg. 11. Consultat el 21 d'agost de 2024 a: https://doi.org/10.1162/DESI_a_00087

s'associa innocent també a coherent o genuí, enlloc d'ingenu). En canvi, el contrari d'una proposició innocent hi ha la proposició mal intencionada (o incoherent), que es donaria quan l'emissor pretén enganyar l'emissor perquè té un interès directe o indirecte en fer-ho i aquest engany pot beneficiar-lo també de forma directa o indirecte (per directe o indirecte es vol apuntar que no sempre una proposició s'utilitza pel que significa en ella mateixa, sinó com una peça de domino que permet generar una contrapartida en un camp aparentment allunyat del contingut de la proposició estudiada). Així mateix, es considera que una afirmació cridanera o estrafolària, volgudament exagerada i provocativa, tot i que pot tenir un impacte social puntual molt elevat en el públic (especialment si s'aprofita dels mitjans de comunicació o disposa d'un fort aparell propagandístic), a llarg termini s'acaba desinflant i el seu grau de perillositat final és més baix que no pas una afirmació raonada i menys provocativa, que, tot i passar desapercebuda, acaba fonamentant una línia d'investigació amb més recorregut. Aquesta consideració sorgeix de l'observació que aquests discursos provocatius requereixen de més repeticions, i els publicats un dia eclipsen els del dia abans, d'aquí la necessitat de fer comunicats tan sovint (com una moda efímer).

Per altra banda, una proposició es considera inofensiva si el receptor no rep cap perjudici fruit de la seva comprensió (ni l'espanta ni el fa enfadar ni, en general, perd possibilitats d'actuació); mentre que es considera perillosa si el receptor, arran d'entendre l'objectiu comunicatiu de la proposició (encara que no necessàriament la proposició en ella mateixa), es veu perjudicat, ja sigui perquè la por el fa prendre males decisions, ja sigui perquè el seu comportament es veu alterat per adequar-se a la realitat plantejada per la proposició en qüestió.

Per exemple, la proposició «L'ordinador s'ha quedat pensant» es considera innocent i inofensiva, ja que tot i que és evident que l'ordinador no pensa sinó que torna uns *outputs* a partir d'uns *inputs* donats i l'execució d'una sèrie d'algoritmes, l'expressió no pot confondre a cap receptor ni fer-lo canviar la seva percepció sobre el món. En canvi, la proposició «LaMDA is a sweet kid who just wants to help the world be a better place»[37] és aparentment genuïna (tanmateix, caldria revisar la biografia completa, especialment, la situació laboral en la que es troba actualment Blake Lemoine després d'haver fet aquesta afirmació), però perillosa, ja que confon al receptor a l'intentar fer-li creure que un LLM és un ésser viu. A l'igual que «L'ordinador és com un cervell», ja que està descrivint com idear un ordinador prenent de model el cervell humà, però si s'arribés a aconseguir, podria tenir un impacte social alt. Per altra banda, es consideren proposicions malintencionades aquelles que s'emeten o bé coneixent que la informació que es dona és falsa, com

37 TIKU, Nitasha (11.06.2022). "The Google engineer who thinks the company's AI has come to life" en *The Washington Post*. Consultat el 13 de juny de 2022 a: https://www.washingtonpost.com/technology/2022/06/11/google-ai-lamda-blake-lemoine/

«La IA eliminarà la humanitat» o bé perquè, encara que no es faci conscientment o de mala fe, tenen un impacte social alt a partir d'una afirmació objectivament falsa, com «El cervell és com un ordinador». En aquest treball s'identifiquen aquest últim tipus de proposicions amb la figura de la inversió de la metàfora computacional, i es defensa que són la base conceptual d'una etologia digital.

Per tant, per il·lustrar aquesta classificació, i exclusivament amb un objectiu explicatiu i sense pretensió quantitativa, es podria fer el següent quadre en el qual hi ha un exemple prototípic de cada tipus de proposició:

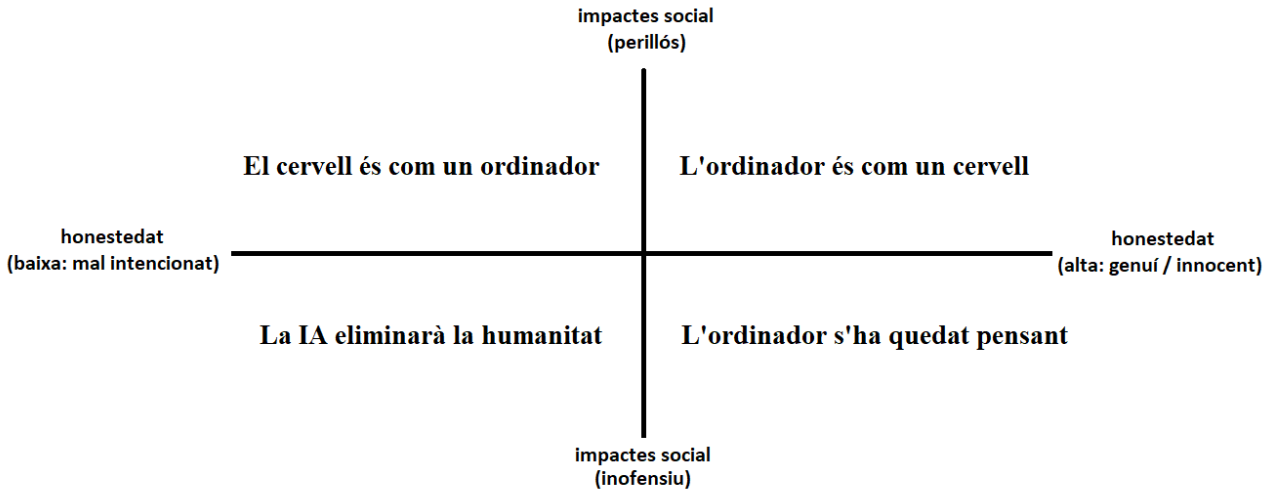

Figura 1: Relació honestedat / impacte social

Així, si es pren l'exemple real següent escrit per Kevin Roose, periodista especialitzat en tecnologia del *New York Times*: «En la conversa de dues hores que vaig tenir amb Bing va revelar una espècie de doble personalitat»[38], tenint en compte 1) el context de la conversa (un article d'opinió en premsa generalista de gran tirada), 2) la intenció comunicativa (entretenir els seus lectors amb una anècdota tecnològica relatada amb dosis d'humor però també cert estupor), 3) el grau de veracitat que pot percebre el receptor (en la mesura que es tracta d'un diari reputat i el periodista és especialista en tecnologia, el lector no ha de per què dubtar de la veracitat del

38 ROOSE, Kevin (25.02.02023). "El xatbot de Bing va dir que m'estimava i pretenia que deixés la meva dona" originalment en *The New York Times* el 16.02.2023 i traduït pel diari *Ara*. Consultat el 8 de juliol de 2023 a: https://www.ara.cat/economia/tecnologia/xatbot-bing-dir-m-estimava-pretenia-deixes-meva-dona_130_4634725.html

contingut) i 4) les conclusions a les que pot arribar (Roose conclou l'article afirmant que, tot i que sap que no són humans, Bing ha realitzat accions com al·lucinar que també fan els humans, experiència que el porta a pensar que s'ha creuat una línia -argument evidentment fal·laç al barrejar el concepte tècnic d'al·lucinar amb el concepte humà d'al·lucinar-). Per tots aquest motius, aquesta proposició es podria situar en aquest punt de la taula.

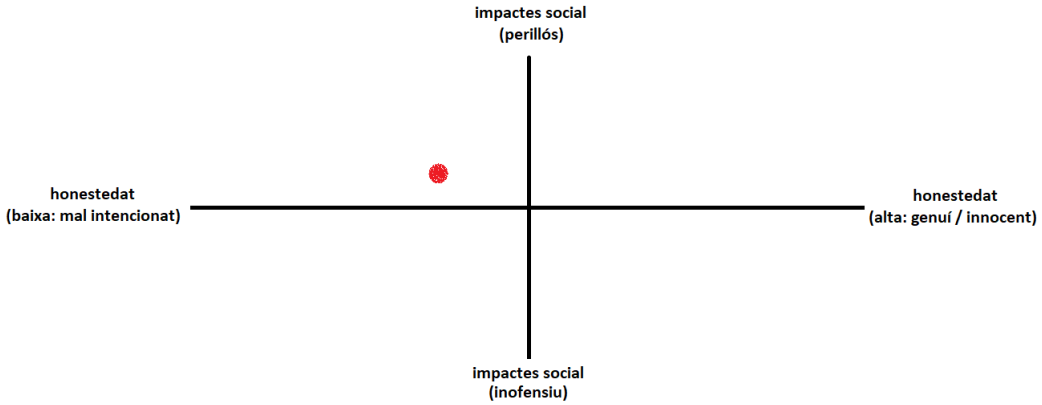

Figura 2: Relació honestedat / impacte social de Kevin Roose

És dir, és tan malintencionada com perillosa (potser més deshonesta que perillosa), ja que l'autor afirma reiteradament que entén com funcionen internament aquests LLM i coneix l'especificitat del lèxic que utilitza i, tot i així, barreja terminologia divulgativa amb termes tècnics i aprofita el seu accés a un públic no necessàriament especialitzat per transmetre una idea falsa i atemorir-lo (de fet, l'entrada biogràfica a la seva web queda clar que viu precisament de fer això[39]). Per tant, aquesta proposició contribueix a la creació d'una etologia digital en la mesura que presenta de forma mal intencionada i perillosa una eina LLM com un ésser viu.

Aquesta taula, com ja s'ha dit abans, no té cap pretensió quantitativa i no hi ha una fórmula matemàtica a partir de la qual es pugui obtenir un resultat. La seva validesa depèn de la raonabilitat dels arguments i proves aportades. Per tant, una nova evidència, com la descoberta que l'emissor d'una proposició ja estudiada ha estat retribuït per una entitat interessada, pot canviar l'avaluació del dictamen.

En els capítols següents s'analitzaran tres tipologies diferents d'etologia digital, des de la més mal intencionada però puerilment perillosa, a la més innocent i només teòricament inofensiva,

39 ROOSE, Kevin. *Kevin Rose. Technology writer*. Consultat el 8 de juliol de 2023 a: https://www.kevinroose.com/bio

passant per una mostra de dues actituds aparentment contraposades que comparteixen un punt intermedi en aquesta classificació.

1.4 Tres actituds i un índex

La primera actitud, la dubtosament honesta o, en alguns moments, obertament malintencionada, la defensen una sèrie d'autors vinculats al concepte de risc existencial i la singularitat, com Stuart Russell, Nick Bostrom o David Chalmers. Són autors que poden ser qualificats de *criti-hypers*[40], ja que, en un moment o altre de la seva vida (doncs les seves posicions també han canviat durant els anys), han fet afirmacions del següent tipus:

- Stuart Russell: «The arrival of superintelligent AI is in many ways analogous to the arrival of a superior alien civilization but much more likely to occur».[41]
- Nick Bostrom: «Recent rapid advances in artificial intelligence makes it timely to start considering what a future society might look like in which humans share the world with digital minds of various kinds and sophistication».[42]
- David Chalmers: «To determine whether we can play a significant role in a post-singularity world, we need to know whether human identity can survive the enhancing of our cognitive systems, perhaps through uploading onto new technology».[43]

S'analitzaran en detall alguns dels seus documents en el capítol 2 d'aquest treball.

Com a punt intermedi es prenen les dues cartes obertes que van escriure Elon Musk i companyia per una banda, i Bill Gates per l'altra. La primera està més a prop del grup anterior, i és interessant en la mesura que contrasta amb la de Gates, més pròxima a la honestedat, però per això potser més perillosa que l'actitud anterior (la bona fe no necessàriament està renyida amb la maldat quan l'objectiu és perillós).

40 Com es veurà en el següent capítol, aquesta expressió la va encunyar Lee Vinsel per designar aquells qui, sota l'aparença d'una crítica, tenen per objectiu generar (falses) expectatives sobre un fenomen.

41 RUSSELL, Stuart (2019). *Human Compatible. AI and the Problem of Control*, Londres, Penguin Books UK, 2020, pàg. 2

42 BOSTROM, Nick; SHULMAN, Carl (2023). "Propositions Concerning Digital Minds and Society" properament en *Cambridge Journal of Law, Politics, and Art*, Issue 3, 2024. Consultat el 5 de juliol de 2024 a: https://nickbostrom.com/propositions.pdf

43 CHALMERS, David (2010). "The Singularity: A Philosophical Analysis" en *Journal of Consciousness Studies* 17:7-65, 2010, pàg.4. Consultat el 5 de juliol de 2024 a: https://consc.net/papers/singularity.pdf

- A la carta de Musk i companyia es troben afirmacions com les següents: «AI systems with human-competitive intelligence can pose profound risks to society and humanity»[44] o «develop and deploy ever more powerful digital minds that no one – not even their creators – can understand, predict, or reliably control».[45]
- A la carta oberta de Gates s'hi poden llegir proposicions com aquestes: «The rise of AI will free people up to do things that software never will—teaching, caring for patients, and supporting the elderly, for example».[46]

S'analitzaran en detall aquests dos documents i d'altres de relacionats en el capítol 3 d'aquest treball.

Com a exemple de la tercera actitud, es prenen l'exemple d'autors com Gary Marcus (i Ernst Davis), Melanie Mitchell o Rodney Brook. Són autors aparentment crítics i, en part, escèptics amb aquest corrent alarmista preponderant (*criti-hype*) i denuncien les seves exageracions, ara bé, defensen la possibilitat d'una autèntica intel·ligència artificial emmirallant-se en la naturalesa i, concretament, en l'evolució de les espècies. Han fet afirmacions com les següents:

- Gary Marcus i Ernst Davies: «If AI could read and reason as well as humans […] science and technology might accelerate rapidly, with huge implications for medicine and the environment and more».[47]
- Melanie Mitchell: «In order to understand the nature of true progress in AI, and in particular, why it is harder than we think, we need to move from alchemy to developing a scientific understanding of intelligence».[48]
- Rodney Brooks: «Calm down people. We neither have super powerful AI around the corner, nor the end of the caused by AI about to come down upon us […]. In short, there will be valuable tools produced, and at the same time lots of damaging misuse».[49]

S'analitzaran en detall algunes de les seves afirmacions en el capítol 4 d'aquest treball.

44 FLI (22.03.2023). "Pause Giant AI Experiments: An Open Letter" en *Futur of Life Institute*, §1. Consultat el 14 de juliol de 2023 a: https://futureoflife.org/open-letter/pause-giant-ai-experiments/

45 FLI (22.03.2023). "Pause Giant AI Experiments: An Open Letter" en *Futur of Life Institute*, §1. Consultat el 16 de juliol de 2023 a: https://futureoflife.org/open-letter/pause-giant-ai-experiments/

46 GATES, Bill (21.03.2023). "The Age of AI has begun" en *GatesNotes. The blog of Bill Gates,* §24. Consultat el 6 de juliol de 2023 a: *https://www.gatesnotes.com/The-Age-of-AI-Has-Begun*

47 MARCUS, Gary; DAVIS, Ernest (2019). *Rebooting AI. Building Artificial Intelligence We Can Trust*, Nova York, Vintage books de Penguin Random House, 2020, pàg. 14.

48 MITCHELL, Melanie (26.04.2021). "Why AI is Harder Than We Think" en *arXiv*, pàg. 8. Consultat el 4 d'agost de 2023 a: https://arxiv.org/abs/2104.12871v1

Finalment, en el capítol que serveix de transició entre la primera part d'aquest treball i la segona, es recorre a una idea de George Zarkadakis, la de les metàfores de cada època, i s'analitza un article de Murray Shanahan, qui defensa que hi ha un problema de llenguatge en la base del projecte de la IA («it takes time for new language to settle, and for new ways of talking to find their place in human affairs. It may require an extensive period of interacting with, of living with, these new kinds of artefact before we learn how best to talk about them»[50]).

Aquest problema de llenguatge, de fet, de la relació entre les idees, les paraules i les coses, és el que s'analitza a la segona part d'aquest treball, quan s'estudia el mal ús de la metàfora computacional, des del seu naixement, no precisament innocu, així com la seva mala interpretació en quelcom que aquí s'ha anomenat la inversió de la metàfora computacional. La hipòtesi és que és aquesta figura conceptual la que permet la confecció d'una etologia digital.

49 BROOKS, Rodney (23.03.2023). "What Will Transformers Transform?" en *Robots, AI, and other stuff*. Consultat el 5 de juliol de 2024 a: https://rodneybrooks.com/what-will-transformers-transform/

50 SHANAHAN, Murray (16.02.2023). "Talking About Large Language Models" en *arXiv* preprint arXiv:2212.03551, pàg. 11. Consultat el 8 d'abril de 2023 a: https://arxiv.org/pdf/2212.03551.pdf

Al decir todo esto me veo ahora inmerso en el universo
de las tecno-imágenes y no en su orilla, como lo hice
hasta ahora en este ensayo. Desde esta posición, puedo
elogiar la superficie y la superficialidad.

Vilém Flusser, *El universo de las imágenes técnicas*

2. Els autors de la por

Amb l'objectiu de presentar evidències de l'existència d'un discurs etològic digital, s'ha confeccionat una mena de mostrari no orientat ni exhaustiu de diferents autors i textos, agrupant-los segons l'estratègia utilitzada. En aquest capítol s'analitzarà el discurs de tres autors representatius d'un tipus de discurs basat en la por. A continuació es justifica aquesta denominació i la tria dels autors.

2.1 L'estratègia de la por i la seva legitimitat

Segons els neurobiòlegs, la por és un mecanisme d'autoprotecció que evolutivament ha servit per arribar fins aquí: sense la por, difícilment molts espècimens arribarien a l'edat suficient per reproduir-se[51]. Aquesta por genètica també pot esdevenir un gran element de motivació, especialment entre aquells que volen que les seves empreses segueixin creixent i multiplicant-se: si hom no vol ser devorat per la competència, ha de recórrer a les armes que aquesta, de ben segur, aconseguirà: «If companies believe a labor-saving technology is so powerful or efficient that their competitors are sure to adopt it, they don't want to miss out — regardless of the ultimate utility»[52]. Per tant, forma part de l'estratègia empresarial sobrevalorar una eina tecnològica que traslladi a la possible clientela les gestes extraordinàries que el seu producte els permetrà, independentment de la necessitat d'aquestes gestes i, encara menys, de la seva bondat. I és també aquesta mateixa estratègia empresarial la que fa que, enlloc de retransmetre a través dels canals comercials les bondats (i maldats) d'aquest producte, canals que generen poca confiança, es faci a través de canals que són percebuts com més objectius i independents, d'alguna manera, exclosos dels circuits comercials: els investigadors.

51 TIMBERLAKE W, Lucas GA. "Behavior system and learning: from misbehavior to general principles". Citat en MISSLIN, René (27.01.2003). "The defense system of fear: behavior and neurocircuitry" en *Neurophysiologie clinique* 33 (2003), pàgs. 55–66.

52 MERCHANT, Brian (31.03.2023). "Afraid of AI? The startups selling it want you to be" en *Los Angeles Times*. Consultat el 2 d'agost de 2023 a: https://www.latimes.com/business/technology/story/2023-03-31/column-afraid-of-ai-the-startups-selling-it-want-you-to-be

En aquest context, Lee Vinsel va crear un terme específic, *criti-hype*, per designar els autors que, sota una aparent crítica cap a una nova tecnologia, el que realment fan és potenciar-la desmesuradament: «I've become increasingly aware of critical writing that is parasitic upon and even inflates hype»[53]. Pel que sembla, una predicció apocalíptica resulta més fàcil de transmetre i en l'era dels pesca-clics això constitueix un avantatge per si mateix. Si aquesta predicció apocalíptica va acompanyada d'un enemic clarament identificat, amb rostre i atributs malèfics, encara és més fàcil. Així que el camp dels discursos apocalíptics sobre la IA és especialment fèrtil per fer créixer etologies digitals.

Entre aquests autors, se n'han triat tres com a exemple d'aquesta tendència. Els tres tenen un bona reputació investigadora i els tres han dedicat bona part de la seva vida al tema de la IA, tot i que els interessos particulars de cada un d'ells no sempre són els mateixos ni sempre són clars. Els autors i les obres a estudiar són els següents:

- Stuart Russell, *Human Compatible* i "Human-Compatible": l'estudi d'aquests dos documents serveix, no només per posar exemples de proposicions que ajuden a confeccionar una etologia digital, sinó també per descartar que aquesta estratègia de la por es faci només com a recurs divulgatiu; per això, se selecciona un text divulgatiu com el primer, i la seva versió acadèmica, com el segon.
- Nick Bostrom, "Sharing the World with Digital Minds" i "Propositions Concerning Digital Minds and Society": els textos estudiats, ambdós documents acadèmics, són dos exemples més d'aquesta estratègia de la por, però mentre el primer confecciona una etologia digital, el segon fa un pas més i comença a preparar el terreny per a una mitologia digital.
- David Chalmers, "Could a Large Language Model be Conscious?": en aquest text es pot veure com també dins d'aquesta estratègia i tot i coincidir amb algunes proposicions amb el discurs de la por, una proposta més honesta pot encara ser més perillosament fecunda per la confecció d'una etologia digital.

Ara bé, no tothom és partidari d'aquesta estratègia, hi ha autors –que alguns pensen que estan més silenciats i d'altres que simplement diuen que són menys comercials– que denuncien aquests tipus de discursos per innecessaris i, sobretot, il·legítims. Una de les denúncies habituals és que, en alguns casos, aquests discursos amaguen interessos econòmics, per això en el seu anàlisi també es posarà sobre la taula aquesta possible variable. Altres denuncien que l'opacitat de les seves

53 VINSEL, Lee (01.02.2023). "You're Doing It Wrong: Notes on Criticism and Technology Hype" en *Medium*. Consultat el 27 de juliol de 2023 a: https://sts-news.medium.com/youre-doing-it-wrong-notes-on-criticism-and-technology-hype-18b08b4307e5

intencions és similar a la dels polítics, cosa que fa molt difícil acabar d'entendre quin interès real tenen en la qüestió: «Like politicians, one cannot simply and naively assume that these people are being honest about their views, wishes, and what they would do. [...] some people seem super-strategic and willing to say whatever will achieve their goals, regardless of whether they believe the claims they make»[54]. En aquest sentit, altres investigadors en ètica de la IA també apunten a la cortina de fum que aquests discursos poden suposar per amagar problemes reals: «Timnit Gebru and Meredith Whittaker have been shouting into the void that an abstract fear of an imminent SkyNet misses the forest for the trees»[55]. Així doncs, allà on es vegi necessari i possible, es contrastarà les paraules d'aquests textos amb les crítiques que han rebut.

2.2 Stuart Russell, *Human Compatible*

Stuart Russell (1962) és professor de Ciència de la Computació i titular de la Càtedra Smith-Zadeh d'Enginyeria a la Universitat de Califòrnia, Berkeley. Ha exercit com a vicepresident del Consell d'IA i Robòtica del Fòrum Econòmic Mundial i com a conseller de les Nacions Unides en matèria de control d'armes[56]. El seu llibre *Artificial Intelligence: A Modern Approach* (amb Peter Norvig) és un text de referència en intel·ligència artificial; ha estat traduït a 14 idiomes i s'utilitza en 1500 universitats de 135 països. La seva recerca abasta una àmplia gamma de temes en intel·ligència artificial, incloent-hi aprenentatge automàtic, raonament probabilístic, representació del coneixement, planificació, presa de decisions en temps real, seguiment de múltiples objectius, visió per ordinador, fisiologia computacional i fonaments filosòfics[57].

L'obra titulada *Human Compatible. AI and the Problem of Control* va ser publicada el 2019 per Vicking als EUA i per Allen Lane al Regne Unit, i a partir de 2020 la publica Penguin Random House. Va rebre tant crítiques positives –per part de premsa generalista o no especialitzada en la matèria com *The Guardian*[58] o *Financial Times*[59] o el *Wall Street Journal*[60]–, com crítiques negatives per part, per exemple, de David Leslie, professor d'ètica a l'Alan Turing Institute –qui el va acusar

54 TORRES, Émilie P. (28.07.2021). "The Dangerous Ideas of «Longtermism» and «Existential Risk»" en *Current Affairs*. Consultat el 2 d'agost de 2023 a: https://www.currentaffairs.org/2021/07/the-dangerous-ideas-of-longtermism-and-existential-risk

55 MERCHANT, Brian (31.03.2023). "Afraid of AI? The startups selling it want you to be" en *Los Angeles Times*. Consultat el 2 d'agost de 2023 a: https://www.latimes.com/business/technology/story/2023-03-31/column-afraid-of-ai-the-startups-selling-it-want-you-to-be

56 "Stuart Russell". A *Penguin Random House*. Consultat el 17 de juliol de 2023 a: https://www.penguinrandomhouse.com/authors/2159461/stuart-russell

57 "Stuart Russell – Biography", "People @EECS" a *Berkeley Electrical Engineering and Computer Sciences*. Consultat el 18 de juliol de 2023 a: http://people.eecs.berkeley.edu/~russell/biography.html

d'amagar els veritables problemes de control sobre la IA sota l'estora de l'arribada d'unes criatures super-intel·ligents[61]–, o Melanie Mitchell, autora d'*Artificial Intelligence: A Guide for Thinking Humans* i professora al Santa Fe Institute –qui assenyalava que era difícil d'imaginar com una super-intel·ligència podia tenir més habilitats socials que els humans sense patir cap restricció de les seves capacitats mecàniques[62].

El llibre té deu capítols i quatre apèndixs. En el primer capítol, titulat "If we succeed", Russell utilitza, com a mínim, les següents expressions etològiques:

P1 «The arrival of superintelligent AI is in many ways analogous to the arrival of a superior alien civilization but much more likely to occur».[63]

P2 «Like any rational entity, the algorithm learns how to modify the state of its environment – in this case, the user's mind– in order to maximize its own reward».[64]

P3 «Machines are **beneficial** to the extent that **their** actions can be expected to achieve **our** objectives».[65]

P4 «The result will be a new relationship between humans and machines, one that I hope will enable us to navigate the next few decades successfully».[66]

Si s'agrupen per tipus de mecanisme etològic, es pot observar com P1 i P4 tracten la IA com una espècie invasora amb la qual no queda més remei que conviure. A P2 se li atribueixen

58 SAMPLE, Ian (24.10.2019). "Human Compatible by Stuart Russell review – AI and our future" a *The Guardian*. Consultat el 18 de juliol de 2023 a: https://www.theguardian.com/books/2019/oct/24/human-compatible-ai-problem-control-stuart-russell-review

59 WATERS, Richard (18.10.2019). "Human Compatible — can we keep control over a superintelligence?" a *Financial Times*. Consultat el 18 de juliol de 2023 a: https://www.ft.com/content/0e79832c-ef48-11e9-bfa4-b25f11f42901

60 HUTSON, Matthew (19.11.2019). "'Human Compatible' and 'Artificial Intelligence' Review: Learn Like a Machine" a *The Wall Street Journal*. Consultat el 18 de juliol de 2023 a: https://www.wsj.com/articles/human-compatible-and-artificial-intelligence-review-learn-like-a-machine-11574207170

61 LESLIE, David (02.10.2019). "Raging robots, hapless humans: the AI dystopia" a *Nature*. Consultat el 18 de juliol de 2023 a: https://www.nature.com/articles/d41586-019-02939-0

62 MITCHELL, Melanie (31.10.2019). "We Shouldn't be Scared by 'Superintelligent A.I.'" a *New York Times*. Consultat el 18 de juliol de 2023 a: https://www.nytimes.com/2019/10/31/opinion/superintelligent-artificial-intelligence.html

63 RUSSELL, Stuart (2019). *Human Compatible. AI and the Problem of Control*, Londres, Penguin Books UK, 2020, pàg. 2.

64 *Ibídem*, pàg. 9.

65 *Ibídem*, pàg. 11. Les negretes són de Russell.

66 *Ibídem*, pàg. 12.

propietats dels humans com la racionalitat i la capacitat d'aprenentatge, així com s'utilitza un terme originalment biològic ("environment") i un altre originalment psicològico-policial ("reward"), termes que, tanmateix, ja han passat a ser comuns també en el món digital. A P3, en el context d'explicar l'autonomia de la IA, sembla que els hi retira aquesta possibilitat, cosa que converteix la proposició o bé en una obvietat que no caldria escriure (si se substitueix "machines" pel terme "martell" la frase sona ridícula i tot), o bé d'una crueltat immensa (si es substitueix "machines" pel terme "toros" la frase sona taurista). Per tant, des d'aquesta aproximació, la IA seria una nova espècie invasora de la qual cal protegir-se capant les seves potencialitats, afirmació que trobaríem èticament acceptable si l'apliquéssim a una planta, però no a un animal amb intel·ligència, que és el que aquí Russell sembla està descrivint. Totes aquestes estratègies corresponen a les identificades com intuïtives.

En el segon capítol, titulat "Intelligence in humans and machines", Russell presenta la justificació del cos teòric de la seva proposta, és a dir, per què cal que els objectius de la IA estiguin alineats amb els objectius dels humans (*the alignment problem*, ho solen anomenar). El capítol, el més llarg del llibre (49 pàgines), es pot separar en dues parts: una part més conceptual, en la qual Russell utilitza l'evolució com a concepte clau; i una part més tècnica, en la qual explica el tractament de la incertesa a nivell computacional. Hi ha diferents mostres d'etologia digital en la primera part, així com també una implícita visió finalista de l'evolució que, en alguns moments, sembla basada en un disseny intel·ligent, comportaments que en el primer capítol d'aquest treball s'han identificat amb una etologia digital raonada. En canvi, en la segona part hi ha expressions que, no només no constitueixen una etologia digital sinó que, fins i tot, la desacrediten. Aquest tipus de contrastos fan que sigui difícil d'entendre l'objectiu de Russell per escriure així aquest llibre; per intentar-ho, s'estudia també un article acadèmic que Russell utilitza per sintetitzar les idees principals del llibre.

Els exemples d'etologia digital del segon capítol són les següents:

P5 «Yet every step towards an explanation of how the mind works is also a step towards the creation of the mind's capabilities in an artifact– that is, a step toward artificial intelligence».[67]

P6 «[...] an entity is intelligent to the extent that what it does is likely to achieve what it wants, given what it has perceived».[68]

67 *Ibídem*, pàg. 14.

68 *Ídem*.

P7 «Evolution doesn't know, in advance, where the glucose is going to be or where your keys are, so putting the capability to find them into the organism is the next best thing».[69]

P8 «It [*E. coli*] never learns. It has no brain, just a few simple chemical reactions to do the job».[70]

P9 «While slow [els cervell humans] compared to electronic circuits, the "cycle time" of a few milliseconds per stage change is fast compared to most biological processes».[71]

P10 «the neural implementation of the *cognitive* level –learning, knowing, remembering, reasoning, planning, deciding, and so on– is still mostly anyone's guest. (Perhaps that will change as we understand more about AI [...])».[72]

P11 «One reason we understand the brain's reward system is that it resembles the method of reinforcement learning developed in AI, for which we have a very solid theory».[73]

Com es pot llegir a P5 i P11, hi ha una relació entre el funcionament natural i el funcionament artificial. Mentre que P5 es pot pensar que això voldria dir que les creacions tecnològiques pretenen ser còpies de la naturalesa (com Da Vinci va estudiar els ocells abans de posar-se a dissenyar una màquina voladora[74]), aquesta interpretació queda rebatuda implícitament per P9, en què es compara la velocitat inferior del cervell amb la velocitat de circuits elèctrics, i explícitament en P11, quan es reconeix que, de fet, és a través dels conceptes artificials que cal entendre la naturalesa: els humans creem eines/mètodes que després observem en la naturalesa. Per tant, es passa d'utilitzar elements intuïtius d'una etologia digital (com l'ús de metàfores), amb poc impacte social i, per si sols, prou honestos, a una inversió de la metàfora computacional, mecanisme tipificat com el més perillós i mal intencionat en la construcció d'una etologia digital, ja que inclou elements intuïtius i altres de raonats (com l'analogia amb la teoria de l'evolució). A més a més, és especialment perillosa perquè la seva estratègia consisteix en minoritzar l'espècie humana i fer-la deutora de la creació digital, és a dir, no és l'ordinador que funciona com el cervell (ja que el seu disseny ha estat basat en allò que se sap del cervell), sinó que és el cervell el que funciona com l'ordinador. I tot allò que la teoria

69 *Ibídem*, pàg. 15.

70 *Ídem*.

71 *Ibídem*, pàg. 16.

72 *Ídem.*

73 *Ibídem*, pàg. 17.

74 CERVERÓ MELIÁ, Ernesto; FERRER GISBERT, Pablo; CAPUZ RIZO, Salvador (11-13.07.2018). "THE DESIGN BASED ON ANALOGIES IN THE WORK OF LEONARDO DA VINCI" en *22nd International Congress on Project Management and Engineering*. Consultat el 19 de juliol de 2023 a: http://dspace.aeipro.com/xmlui/handle/123456789/1624

computacional no pugui descriure (que és on hi ha teories sòlides) passa a ser un mal funcionalment natural. Es tractarà més a fons la metàfora computacional i la seva inversió a la segona part d'aquest treball.

Per arribar a aquesta inversió, Russell ha partit d'una jerarquització de la naturalesa més o menys explícita: per P8, els bacteris, els quals no tenen cervell, actuen per reacció i no tenen un mecanisme d'aprenentatge; per P7, aquest mecanisme d'aprenentatge *ve posat* per l'evolució (talment un agent intencional), i permet identificar éssers amb intel·ligència, definida en P6 com el mecanisme d'aconseguir el que es vol (cosa que implica que o bé els bacteris, tot i que no aprenen i no tenen cervell, són intel·ligents, ja que aconsegueixen la glucosa, o bé l'evolució és intel·ligent i ha dissenyat un sistema sense aprenentatge i sense cervell, però que permet sobreviure als bacteris). I per P9, els humans, tot i que sí que tenim cervell, aquest no és tan ràpid com mecanismes que funcionen amb circuits elèctrics, això és, els ordinadors. Per tant, en una estratègia bastant comú en una etologia digital, Russell estableix una jerarquia per comparació, jerarquia a la part baixa de la qual hi ha els bacteris i a la part alta els ordinadors (els humans semblen estar una mica per sota).

Per aconseguir construir aquesta jerarquia, Russell ha partit d'un element a comparar (l'aprenentatge) i d'una escala sobre la que comparar (l'evolució).

P12 «Learning is good for more than surviving and prospering. It also *speeds up evolution*».[75]

P13 «After all, learning doesn't change one's DNA, and evolution is all about changing DNA over generations».[76]

P14 «Clearly, if evolution has to worry about choosing only the first three digits, its job is much easier; the adaptive organism, in learning the last three digits, is doing in one lifetime what evolution would have taken many generations to do».[77]

P15 «"How did the reward system get there in the first place?" The answer, of course, is by an evolutionary process, one that internalized a feedback mechanism that is at least somewhat aligned with evolutionary fitness».[78]

P16 «Evolution considers you only as an agent, that is, something that acts».[79]

P17 «One reason artificial intelligence is so fascinating is that it offers a potential route to understanding these issues: we may come to understand both how these intellectual

75 RUSSELL, Stuart (2019). *Human Compatible. AI and the Problem of Control*, Londres, Penguin Books UK, 2020, pàg. 18.

76 *Ídem.*

77 *Ibídem*, pàg. 19.

78 *Ídem*.

79 *Ídem*.

characteristics make intelligent behavior possible and why it's impossible to produce truly intelligent behavior without them».[80]

Com es pot comprovar, l'ús del verb *aprendre* es fa servir en diferents sentits: des d'un sentit biològic com en P13, a un sentit tecnològic com en P11, a un més comú i proper a la pedagogia com en P14. Aparentment, el que es compara són diferents mecanismes d'aprenentatge, entre els quals hi ha des d'aquells que es poden atribuir a un bacteri, als que es poden atribuir a un infant, o els que es poden programar perquè un algoritme modifiqui un *output* per un altre. Aquesta comparació se sustenta en què en qualsevol dels tres sentits d'aprenentatge hi intervé el que en psicologia s'anomena un sistema de recompensa. Ara bé, ni és el mateix l'aprenentatge en un sentit evolutiu que pedagògic, ni és el mateix un sistema de recompensa neuronal que un sistema de recompensa computacional (ni tampoc en sentit penal, que és d'on sorgeix l'expressió[81]). En aquest sentit, aquest tipus de transmissió semàntica és molt semblant a la que es fa en etologia, tal i com explica Frans de Waal:

> La ciència no persegueix comprendre el fetge de la rata o el fetge humà, sinó el fetge, i punt. Tots els òrgans i processos biològics són molt més antics que la nostra espècies, i han evolucionat durant milions d'anys només amb unes poques modificacions específiques de cada organisme. Així és com funciona sempre l'evolució. Per què hauria de ser diferent la cognició? La nostra primera tasca és esbrinar com opera la cognició en general, quins elements requereix la seva funció, i com se sintonitzen aquest elements amb els sistemes sensorials i l'ecologia de cada espècie.[82]

En el cas de l'aprenentatge, el sistema de recompensa seria una funció d'un *òrgan* que permetria fer el pont per descobrir el funcionament general de la cognició. Ara bé, per a de Waal es tracta de comparar un òrgan donat (com el fetge d'una rata) amb un altre òrgan donat (com el fetge humà) per entendre la funció del fetge en general (i, si es vol, construir un fetge artificial per persones amb cirrosis), però no tindria massa sentit després prendre el fetge artificial per arribar a conclusions sobre el fetge de la rata, atribuint-li aquest algun tipus de defecte en el disseny. De fet, aquesta segona interpretació, la de Russell, acaba de prendre sentit si es té en compte un paràgraf anterior en el qual, amb la intenció de descartar la necessitat d'entendre la consciència, Russell

80 *Ibídem*, pàgs. 19-20.

81 Reward (n). A *Online Etymology Dictionary*: Consultat el 22 de juliol de 2023 a: https://www.etymonline.com/word/reward

82 DE WAAL, Frans (2016), *¿Tenemos suficiente inteligencia para entender la inteligencia de los animales?*, Barcelona, Tusquets Editores, 2019, pàg. 185. Traducció pròpia.

defensa que es pot saber com funciona l'aprenentatge simplement a través de la conducta del subjecte:

> Suppose I give you a program and ask, "Does this present a threat to humanity?" You analyze the code and indeed, when run, the code will form and carry out a plan whose result will be the destruction of the human race, just as a chess program will form and carry out a plan whose result will be the defeat of any human who faces it. Now suppose I tell you that the code, when run, also creates a form of machine consciousness. Will that change your prediction? Not at all. It makes absolutely no difference. Your prediction about its behavior is exactly the same, because the prediction is based on the code.[83]

Tot i que l'objectiu de l'argument no és fer l'equiparació entre el codi genètic i el codi de programació, sinó eliminar la consciència com atribut necessari per definir la intel·ligència, Russell necessita que s'accepti aquesta equiparació, ja que és conscient, valgui la redundància, que de la consciència no se'n sap res: «In the area of consciousness, we really do know nothing, so I'm going to say nothing. No one in AI is working on making machines conscious, nor would anyone know where to start, and no behavior has consciousness as a prerequisite»[84] (cosa que és falsa, ja que sí que s'hi està treballant, com es veurà més endavant quan s'analitzi Bostrom i Chalmers). Sense la consciència, el problema de la intel·ligència queda reduït a un estudi del comportament i, en aquest cas, de la predicció del comportament. Si el fet d'utilitzar el terme *code* per referir-se als algoritmes no és casual –i no sembla casual quan l'argument està centrat en si cal o no estudiar la consciència humana per entendre en què consisteix la intel·ligència–, llavors s'està fent una equiparació entre el codi genètic i el codi de programació.

Tornant a l'anàlisi de les proposicions, finalment, en P14 i P16 s'expliciten ara alguna de les implicacions d'aquesta hipòtesi que es veuen recolzades per una interpretació finalista de l'evolució, que és mencionada com un agent intencional que es preocupa (*has to worry*) i considera (*considers*). Es pot argumentar que aquest ús es merament divulgatiu i pretén que un lector no especialitzat entengui alguns conceptes a partir d'una personalització, tanmateix, és sospitós que aquest ús aparegui en un context en què s'ha fet una relació prou evident entre el codi genètic i el codi de programació i en el qual s'ha admès, com es fa en P17, que hi ha un vincle entre els descobriments que es puguin fer en el camp de la IA i els que es puguin fer en el camp de la neurociència. De fet, aquest vincle pretén ser un isomorfisme entre dos conjunts, el conjunt de les teories de la intel·ligència artificial i el conjunt de les teories de la intel·ligència natural

83 RUSSELL, Stuart (2019). *Human Compatible. AI and the Problem of Control*, Londres, Penguin Books UK, 2020, pàgs. 16-17.

84 *Ibídem*, pàg. 16.

(anteriorment, en P11 ja s'ha vist que, fins i tot, és la validesa d'una teoria sobre IA la que permet el descobriment d'un procés neuronal). Si això és així, i es té en compte que el codi de programació està escrit amb una certa intencionalitat, és evident que aquesta intencionalitat s'està implicant també en el codi genètic, cosa que donaria sentit als usos intencionals que es prediquen de l'evolució (i no només com a pràctica divulgativa). En aquest cas, es podria concloure que el fet de desenvolupar una etologia digital té com a conseqüència fonamentar una teoria sobre el disseny intel·ligent.

Ara bé, com s'ha dit al principi d'aquest segon capítol, Russell no es manté coherent en aquesta posició i, en la part més tècnica d'aquest capítol, explícitament descarta alguns aspectes d'aquesta interpretació, en concret, la possibilitat de fer comparacions entre ordinadors i cervells:

P18 «Although comparisons between computers and brains are not especially meaningful, the numbers for Summit slightly exceed the raw capacity of the human brain[...]».[85]

P19 «Even in the 1950s, computers were described in the popular press as "super-brains" that were "faster than Einstein." So can we say now, finally, that computers are as powerful as the human brain? No. Focusing on raw computing power misses the point entirely».[86]

P20 «Contrary to common interpretations, I doubt that the test was intended as a true definition of intelligence, in the sense that a machine is intelligent if and only if it passes the Turing test».[87]

P21 «These sharp boundaries on machine competence mean that when people talk about "machine IQ" increasing rapidly and threatening to exceed human IQ, they are talking nonsense».[88]

P22 «Trying to assign an IQ to machines is like trying to get four-legged animals to compete in a human decathlon. True, horses can run fast and jump high, but they have a lot of trouble with pole-vaulting and throwing the discus».[89]

Sembla, doncs, que l'estratègia seguida pel propi Russell en P9 ara és descartada per P18, P19, P21 i P22 explícitament. Cal preguntar-se, doncs, per què ho ha fet i per què ho tornarà a fer més endavant, per exemple, aquí:

85 *Ibídem*, pàg. 34. Summit és un supercomputador d'IBM que s'utilitza al laboratori d'Oak Ridge, a Tennessee.

86 *Ibídem*, pàg. 36.

87 *Ibídem*, pàg. 41.

88 *Ibídem*, pàg. 48.

89 *Ídem.*

P23 «RL algorithms learn from direct experience of reward signals in the environment, much as a baby learns to stand up from the positive reward of being upright and the negative reward of falling over».[90]

Per acabar d'embolicar la troca, un any abans de la publicació d'aquesta llibre, un usuari identificat com Stuart Russell havia fet la següent manifestació en un comentari d'un blog sobre sistemes de recompensa que funcionen malament en jocs que utilitzen IA:

> The notion of "gaming" and "hack" suggests the AI system knows the user's intent but decides to violate it anyway by sticking to the letter of the objective function. I think that this is likely to be misleading for the lay person. Instead, we should think of these as errors in specifying the objective, period.[91]

Aquest tipus d'afirmacions, que semblen que van en la línia d'aquest treball, és a dir, que reclamen certa cura en l'ús del llenguatge per tal de no transmetre una interpretació errònia d'un enunciat sobre IA, són precisament les que després ell mateix potencia. Tanmateix, els motius que tingui Russell per escriure una cosa ara i més endavant negar-la, no són objecte d'estudi (i sempre es pot suposar que hi ha motius merament estilístics, divulgatius i comercials). Per una banda, es reconeix que l'objectiu és substituir els humans, per exemple, així: «That's the ultimate goal of AI research: a system that needs no problem-specific engineering and can simply be asked to teach a molecular biology class or run a government»[92]. I, per altra banda, Russell es presenta amb una possible solució, que anomena IA beneficiosa, com s'ha vist en P3. Per tant, no es pot negar que aquest tipus de discurs col·labora en generar certa confusió fecunda al voltant d'una etologia digital i també pot posar en dubte el nivell d'honestedat, o, com a mínim, d'ingenuïtat de l'autor (qui té uns coneixements tècnics profusament demostrats). Per tant, o bé són estratègies publicitàries o bé són incoherents barrija-barreges; qualsevol dels dos casos no permeten considerar-lo un autor completament honest en aquest sentit. Per tant, si l'objectiu era generar dubtes en el lector, queda perfectament aconseguit. Per altra banda, potser aquesta és una estratègia més d'una etologia digital: com que no pot ser afirmada acadèmicament, és transmesa divulgativament.

---

90 *Ibídem*, pàg. 55.

91 RUSSELL, Stuart (02.04.2018). "Specification gaming examples in AI" en *Victoria Krakovna*. Consultat el 22 de juliol de 2023 a: https://vkrakovna.wordpress.com/2018/04/02/specification-gaming-examples-in-ai/. El 28 de gener de 2024, en resposta privada per correu electrònic, l'assistent de Stuart Russell, James Paul "J.P." Gonzales, M.S., confirma que el comentari és de l'autèntic Stuart Russell.

92 RUSSELL, Stuart (2019). *Human Compatible. AI and the Problem of Control*, Londres, Penguin Books UK, 2020, pàg. 46.

Per contrastar aquesta hipòtesi, caldria resseguir ara un text acadèmic de Russell en el qual aquest element divulgatiu no pugui ser esgrimit com a argument. Curiosament, aquest text és un resum tècnic que Russell titula igual que el llibre, "Human-Compatible Artificial Intelligence"[93], en el qual va fer l'última modificació el 09.03.2021 a les 11:07 i que penja en la seva entrada de la pàgina oficial de la Universitat de Califòrnia, Berkeley.

"Human-Compatible", versió acadèmica

A diferència de la versió divulgativa, en aquest text acadèmic Russell és especialment curós en evitar pràcticament qualsevol mecanisme propi d'una etologia digital. La seva cura s'estén fins el punt de que el primer verb psicològic atribuït a un objecte digital el fa escrivint la frase en negatiu i el verb entre cometes: «no self-driving car in existence today "knows" that pedestrians prefer not to be run over»[94] (tot i que és cert que el *today* no exclou la possibilitat que un dia això sigui possible). El text també té cura d'identificar correctament l'objecte sobre el qual tracta usant termes específics enlloc d'ambigus: "machine" (96 vegades) i "robot" (23 vegades). Fins i tot, el terme "agent" (15 vegades) és usat en un sentit tècnic, que semànticament tant pot designar una persona com una entitat digital, en expressions com: «The simplest case of an assistance game involves two agents, one human and the other a robot»[95]. L'únic moment en què es personifica un ens digital se li assigna un nom descriptiu, Robbie el robot, cosa que tampoc es pot considerar un plantejament etològic en aquest context. I només a l'estudiar teories d'altres autors, com Steve Omohundro, es planteja la possibilitat que un entitat intel·ligent (eufemisme per dir ens digital) actuï per preservar la seva pròpia existència («act to preserve its own existence»[96]); tanmateix, en la frase següent aclareix Russell: «This tendency has nothing to do with a self-preservation instinct or any other biological notion; it's just that an entity usually cannot achieve its objectives if it's dead»[97]. Per tant, Russell interpreta aquesta existència en un sentit estrictament tècnic.

Tanmateix, sí que indirectament es poden assenyalar alguns usos ambigus en les següents proposicions:

93 RUSSELL, Stuart (09.03.2021). "Human-Compatible Artificial Intelligence" en *Berkeley Electrical Engineering and Computer Sciences*. Consultat el 26 de juliol de 2023 a: https://people.eecs.berkeley.edu/~russell/papers/mi19book-hcai.pdf

94 *Ibídem*, pàg. 3.

95 *Ibídem*, pàg. 8.

96 *Ibídem*, pàg. 3.

97 *Ídem.*

P24 «Bostrom's estimate that superintelligent AI might arrive within this century is roughly consistent with my own, and both are considerably more conservative than those of the typical AI researcher».[98]

P25 «Indeed, one expects to find qualitatively different phenomena occurring when the robot is much less capable than, roughly as capable as, or much more capable than the human».[99]

En P24, per explicar la construcció d'una IA superintel·ligent, es fa servir el verb "arrive", com si aquesta invenció vingués de fora, talment una invasió, cosa que desvincula un invent del seu inventor, reduint la responsabilitat que se li pot atribuir per haver fet aquella invenció. Tot i així, s'atribueix l'afirmació a un altre autor i simplement es corrobora. En P25, tot i que es compara la capacitat d'una robot amb un humà, la comparació identifica clarament la naturalesa de cada un, evitant així cap indici d'etologia.

Menció a part mereix l'atribució d'aprenentatge a aquest objecte correctament identificat com a màquina o robot, com per exemple en els següents casos:

P26 «after seeing Arthur Samuel's checker-playing program learn to play checkers far better than its creator».[100]

P27 «we hope the machine's intelligence will be applied both to learning our true objectives and to helping us achieve them».[101]

P28 «the robot can learn more about human preferences from the observation of human behavior——a process that is the dual of reinforcement learning, wherein behavior is learned from rewards and punishments».[102]

Mentre que P26 i P27 simplement atribueixen una capacitat humana i, per extensió animal, a una entitat digital, en P28 s'explicita que la base d'atribuir aprenentatge a una màquina és possible si es redueix aquest aprenentatge a una aplicació de recompenses i càstigs com a model de configuració d'atributs. Com s'ha vist, l'atribució indiscriminada a un ens digital de la possibilitat d'aprendre se sustenta en aquesta relació entre teoria de l'aprenentatge i una teoria de la recompensa i el càstig d'origen psicològic que, abans de ser utilitzada extensament per corrents conductistes, encara basculava entre el seu sentit estrictament legal i el seu ús en textos

98 *Ibídem*, pàg. 6.

99 *Ibídem*, pàg. 13.

100 *Ibídem*, pàg. 1.

101 *Ibídem*, pàg. 7.

102 *Ibídem*, pàg. 8.

d'experiments amb infants, en frases de 1908 com la següent: «Each time the infant picked up one of them, say red, she was rewarded by being given a taste of honey, syrup or sugar. When she picked up the other (blue) brick, no reward was given her»[103]. En P28, es pot veure com s'estableix aquesta relació a través del concepte tècnic "reinforcement learning". Aquesta tècnica, tal i com afirma Nils J. Nilsson, s'inspira en el concepte d'estímul reforçat que havia popularitzat Skinner: «Skinner's work did, however, provide the idea of a *reinforcing* stimulus»[104]. També és útil veure com s'explica aquesta relació: «Borrowing terms form psychological learning theory, we can call the win or loss information (or in general the good-result or bad-result information) a "reward" or a "reinforcement," and this style of learning is called "reinforcement learning" or (sometimes) "trial-and-error learning"»[105] (segurament, el nom assaig-i-error era més descriptiu de com funciona realment aquesta tècnica, però no va acabar de quallar). Aquesta vinculació és la que permet a Russell invertir una metàfora de caràcter tecnològic de la següent manera:

> P29 Indeed, it seems likely that our preferences are at least partially formed by a process resembling inverse reinforcement learning, whereby we absorb preferences that explain the behavior of those around us. Such a process would tend to give cultures some degree of autonomy from the otherwise homogenizing effects of our dopaminebased reward system.[106]

En aquest text és el procés humà el que s'assembla al procés tecnològic d'aprenentatge per reforç invers, i no el procés tecnològic el que estigui inspirat en el procés humà. Aquest gir constitueix un exemple d'inversió de la metàfora computacional. Fins i tot, Russell fa una explicació del comportament aparentment sense sentit pràctic de la cultura. És a dir, tot el model conceptual que havia sorgit originalment de la teoria penal, passa primer per la psicologia, que la converteix en base conceptual de la teoria conductista i acaba aplicada en el camp de la tecnologia, just per ser aplicada novament a un àmbit humà. Aquest tipus de mecanismes són els més perillosos d'una etologia digital, ja que explícitament o implícita (depenent del cas), minoren les característiques humanes per tal que aquestes encaixin en les potencialitats tecnològiques. Si es

103 MYERS, Charles S. (1908). *"*Some Observations on the Development of the Colour-sense*"* en *British Journal of Psychology*, Londres, Vol. 2, Iss. 4, (Oct 1, 1908), pàg 353.

104 NILSSON, Nils J. (2010). *The Quest for Artificial Intelligence. A History of Ideas and Achievements*, New York, 2010, Cambridge University Press, pàg. 20. La cursiva és de Nilsson.

105 *Ibídem*, pàg, 415.

106 RUSSELL, Stuart (09.03.2021). "Human-Compatible Artificial Intelligence" en *Berkeley Electrical Engineering and Computer Sciences,* pàg. 17. Consultat el 26 de juliol de 2023 a: https://people.eecs.berkeley.edu/~russell/papers/mi19book-hcai.pdf

porta aquesta estratègia a les últimes conseqüències, llavors apareixen expressions pròpies, no només d'una etologia digital, sinó de certa divinització de la IA:

P30 «We cannot all be Ruler of the Universe. This means that machines must mediate among conflicting preferences——something that philosophers and social scientists have struggled with for millennia».[107]

En aquest cas, davant la impossibilitat que cada persona pugui imposar les seves preferències a la resta, és la IA de la qual és propietari la persona la que podria mediar amb la IA de la persona amb la qual hom entre en conflicte d'interessos per dirimir a quin dels dos propietaris cal cedir la preferència. Així mateix, l'última frase del text posa en condicions d'igualtat els humans i la IA:

P31 «Taking the problem seriously seems likely to yield new ways of thinking about AI, its purpose, and our relationship to it».[108]

Per tant, queda acreditat que, tot i que Russell col·labora especialment en la confecció d'una etologia digital en els seus textos més divulgatius, també ho fa en el tècnics, encara que sigui només quan la qüestió de l'aprenentatge el porta a fer elucubracions. Ara bé, com s'ha vist en P30, aquestes elucubracions són tan o més perilloses que algunes de les afirmacions que fa en el text divulgatiu, doncs deixen escapar la veritable cosmovisió de Russell: per evitar la destrucció dels humans, cal una entesa amb aquest nou ésser per aconseguir algun tipus de relació estable. Per tots aquests motius, i per el gran ressò que té els seu discurs a nivell social, es podria representar la col·laboració de Russell en una etologia digital així:

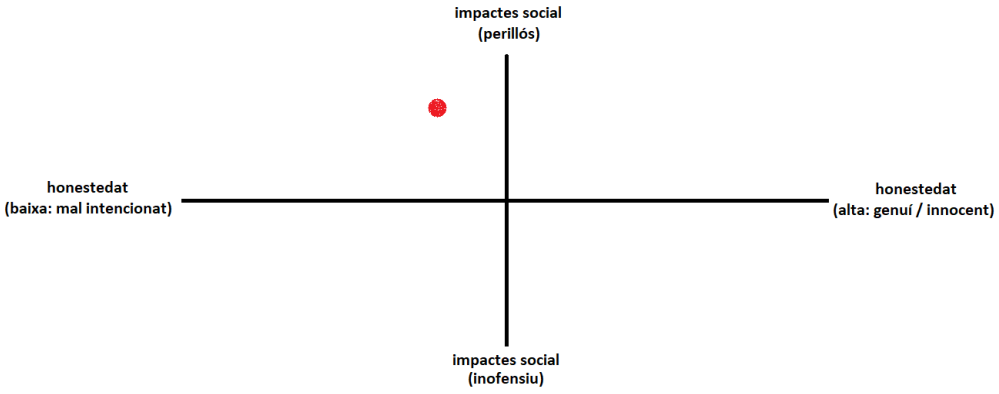

Figura 3: Nivell d'honestedat / impacte social de Stuart Russell

107 *Ibídem*, pàg. 17.

108 *Ibídem*, pàg. 19.

Al gràfic s'intenta il·lustrar la penalització en honestedat (o integritat o incoherència o ingenuïtat), que, tot i no ser severa, pot també fer que el seu impacte social, tot i perillós, a llarg termini sigui menor.

2.3 Nick Bostrom, el supercervell suec

Nick Bostrom (1973) és un filòsof d'origen suec amb formació en física teòrica, neurociència computacional, lògica i intel·ligència artificial, juntament amb filosofia. Segons la seva pròpia autobiografia, és un dels filòsofs més citats del món (30.024 citacions segons Google Scholar[109]) i fa constar en la seva pròpia pàgina que algú s'hi ha referit com «the Swedish superbrain»[110]. Ha estat professor a la Universitat d'Oxford, on va exercir com a director fundador del *Future of Humanity Institute* (FHI) des de 2005 fins al seu tancament a l'abril de 2024. Segons la web de l'entitat, ara només consultable a través de pàgines arxivades, aquest tancament va ser degut a problemes administratius cada vegada més habituals entre aquesta entitat i l'organisme a la qual pertanyia, la Facultat de Filosofia de la Universitat d'Oxford. Segons Bostrom, l'entitat va morir per burocràcia: «I think the death by bureaucracy was regrettable, but there are now so many more places where this [research] can be done»[111]. Tanmateix, altres fonts com *The Guardian*, relacionen el tancament amb algunes conductes poc apropiades:

> The closure of Bostrom's center is a further blow to the effective altruism and long-termism movements that the philosopher had spent decades championing, and which in recent years have become mired in scandals related to racism, sexual harassment and financial fraud. Bostrom himself issued an apology last year after a decades-old email surfaced in which he claimed "Blacks are more stupid than whites" and used the N-word.[112]

De fet, Bostrom és una referència pels directius de Silicon Valley com Elon Musk, Sam Altman o Bill Gates, els quals van fer diverses donacions als seus projectes. És interessant veure la capacitat de Bostrom per acaparar inversions, les més recents de les quals, com mostra en el seu currículum, són les següents:

---

109 "Nick Bostrom" en *Google Scholar*. Consultat el 8 de juliol de 2024 a: https://scholar.google.com/citations?hl=en&user=oQwpz3QAAAAJ&view_op=list_works

110 BOSTROM, Nick. *Nick Bostrom Home's page*. Consultat el 7 de juliol de 2024 a: https://nickbostrom.com/#bio.

111 ADAM, David (26.04.2024). "Future of Humanity Institute shuts: what's next for 'deep future' research?" en *Nature*. Consultat el 7 de juliol de 2024 a: https://www.nature.com/articles/d41586-024-01229-8

112 ROBINS-EARLY, Nick (20.04.2024). "Oxford shuts down institute run by Elon Musk-backed philosopher" en *The Guardian*. Consultat el 7 de juliol de 2024 a: https://www.theguardian.com/technology/2024/apr/19/oxford-future-of-humanity-institute-closes

- £13,200,000. Obtained research grant from the Open Philanthropy Project for the Future of Humanity Institute (2018-2023).
- £1,620,000. Obtained research grant from the Open Philanthropy Project for the Future of Humanity Institute (2017-2026).
- £2,000,000. Obtained research grant from the Leverhulme Trust for the Future of Humanity Institute (2016-2026).
- £1,531,000. Obtained research grant from the European Research Council's Horizon 2020 programme for the Future of Humanity Institute (2015-2020).[113]

Moltes d'aquestes entitats apareixeran de nou quan s'analitzi, en el capítol 3 d'aquest treball, la carta firmada per Elons Musk i companyia. Són entitats finançades pels directius de moltes d'aquestes companyies, com Dustin Moskovitz (cofundador de Facebook) i finançador d'*Open Philanthropy Project*; directius que, per altra banda, no sempre comparteixen estratègies –de fet, Moskovitz va protagonitzar una polèmica amb Musk al denunciar els comptes, segons el seu punt de vista fraudulents, de Tesla[114]–, però que poden compartir inversions i projectes, com també explica un altre article de *The Guardian*:

> Just a month before Bostrom's incendiary comments came to light, the cryptocurrency entrepreneur Sam Bankman-Fried was extradited from the Bahamas to face charges in the US relating to a multibillion-dollar fraud. Bankman-Fried was a vocal and financial supporter of effective altruism and a close friend of William MacAskill, an academic who has strong links to the FHI and who set up the Centre for Effective Altruism, where Bankman-Fried worked briefly.[115]

Aquí, s'analitzaran les proposicions de Bostrom a la recerca d'etologies digitals intentant ignorar aquestes connexions. Tot i que Bostrom va saltar a la fama amb la seva obra *Superintelligence: Paths, Dangers, Strategies* (2014), aquí s'opta per analitzar dos dels articles més

113 BOSTROM, Nick. "Curriculum Vitae" en *Nick Bostrom Home's page*. Consultat el 8 de juliol de 2024 a: https://nickbostrom.com/cv.pdf

114 DANIEL, Will (29.04.2024). "Asana CEO calls Tesla the next Enron and says Elon Musk has misled customers" en *Fortune*. Consultat el 7 de juliol de 2024 a: https://fortune.com/2024/04/29/asana-ceo-tesla-next-enron-elon-musk-misled-customers-investors/

115 ANTHONY, Andrew (28.04.2024). "'Eugenics on steroids': the toxic and contested legacy of Oxford's Future of Humanity Institute" en *The Guardian*. Consultat el 5 de juliol de 2024 a: https://www.theguardian.com/technology/2024/apr/28/nick-bostrom-controversial-future-of-humanity-institute-closure-longtermism-affective-altruism

recents que firma amb Carl Shulman, investigador associat al *Future for Humanity Institute* i assessor a *Open Philanthropy*, qui, per la diferència d'edat, biografia, coincidències laborals i jerarquia sembla pròpiament un dels seus deixebles[116]. El primer, de 2022, titulat "Sharing the World with Digital Minds", conté afirmacions que els autors accepten com a pròpies, mentre que el segon, "Propositions Concerning Digital Minds and Society", escrit el 2023 i pendent de publicació[117], els autors es desmarquen explícitament d'un possible alineament amb les proposicions que escriuen: «We are not ready, at this point, to confidently or "officially" endorse them, nor do they give a full picture of our views on these matters; but we put them forward to facilitate feedback and to invite broader discussion»[118]. Aquesta falca els permet desbravar-se més lliurament, cosa que fa que el text sigui més fecund per a una etologia digital.

Bostrom en "Sharing the World with Digital Minds" i altres proposicions

L'*Abstract* del document comença amb una afirmació que assumeix directament que és qüestió de temps que hi hagi entitats digitals amb ment (*digital minds*):

P32 «The minds of biological creatures occupy a small corner of a much larger space of possible minds that could be created once we master the technology of artificial intelligence».[119]

Assumir que hi ha altres ments que les humanes i, en aquest cas, relacionar el concepte de ment amb la IA, és una clara mostra d'etologia digital. I no només utilitzen aquest recurs, sinó que

116 El dia 7 de juliol de 2024, es pregunta per correu electrònic a l'oficina de Nick Bostrom sobre el tema, apel·lant a la diferència d'ordre en l'autoria dels dos documents, ja que en el primer, l'orde és Shulman i després Bostrom, mentre que en el segon, l'ordre és Bostrom i després Shulman. També s'ha enviat la mateixa pregunta per Likedin a Shulman. El 10 de juliol de 2024, Emily Campbell, Executive Assistant de Nick Bostrom, contesta de part de Bostrom que l'autoria és compartida amb Shulman a parts iguals en els dos articles.

117 Jack Graveney, editor en cap del *Cambridge Journal of Law, Politics, and Art*, revista en la qual ha de sortir publicat l'article en qüestió, ha confirmat, a través d'un correu electrònic rebut l'11 de juliol de 2024, que, tot i que hi hagut un petit retard en la publicació fins a 2025, l'article de Bostrom i Shulman hi apareixerà. En aquest treball es continua referenciant tal i com Bostrom ho especifica en l'article, tot i que ja no es publiqui el 2024.

118 BOSTROM, Nick; SHULMAN, Carl (2023). "Propositions Concerning Digital Minds and Society" properament en *Cambridge Journal of Law, Politics, and Art*, Issue 3, 2024, pàg.1. Consultat el 7 de juliol de 2024 a: https://nickbostrom.com/propositions.pdf

119 SHULMAN, Carl; BOSTROM, Nick (2021). "Sharing the World with Digital Minds" en *Rethinking Moral Status,* Clarke, S., Zohny, H. & Savulescu, J. (eds.), Oxford, Oxford University Press, 2021, Abstract. Consultat el 8 de juliol de 2024 a: https://academic.oup.com/book/41245/chapter/350760172

en una estratègia clàssica de la por, afirmen que el problema sorgirà quan hi hagi una disputa pels limitats recursos naturals:

P33 «Here we focus on one set of issues, which arise from the prospect of digital minds with superhumanly strong claims to resources and influence».[120]

Emfatitzar aquesta capacitat superior a la humana (*superhumanly*) fa evident que la batalla contra aquestes ments digitals està perduda, tret de que els humans ens avancem: és una qüestió d'ells o nosaltres.

P34 «Such beings [individual digital minds with superhuman moral status] could contribute immense value to the world, and failing to respect their interests could produce a moral catastrophe, while a naive way of respecting them could be disastrous for humanity».[121]

Per tant, és raonable que es planifiqui bé aquesta relació amb aquest tipus d'entitats, altrament podia ser catastròfic per a la humanitat: l'apel·lació a la catàstrofe és, per definició, una estratègia de la por, reforç per una etologia digital. Seguidament, en la introducció, els autors plantegen aquesta gran catàstrofe com una gran oportunitat, si és tenen en compte les limitacions humanes i que un bon repartiment podria arribar a ser favorable pels humans si circumstancialment aquestes ments digitals provinguessin d'algun tipus de perfeccionament humà (en una nota a peu de pàgina, es menciona explícitament la idea de David Chalmers de pujar la consciència al núvol):

P35 «Human biological nature imposes many practical limits on what can be done to promote somebody's welfare. We can only live so long, feel so much joy, have so many children, and benefit so much from additional support and resources».[122]

P36 «However, these constraints may loosen for other beings. Consider the possibility of machine minds with conscious experiences, desires, and capacity for reasoning and autonomous decision-making».[123]

P37 «This could be a wonderful development: lives free of pain and disease, bubbling over with happiness, enriched with superhuman awareness and understanding and all manner of higher goods».[124]

120 *Ídem*.

121 *Ídem.*

122 *Ibídem*, pàg. 306.

123 *Ídem.*

124 *Ídem*.

Per tant, ja sigui com a defensa preventiva (P35), ja sigui perquè potser aquestes ments digitals som nosaltres (P37), és a dir, per interès, és una estratègia guanyadora (*win-win*) prendre's seriosament aquesta hipòtesi.

Després d'aquest aclariment, els autors relacionen la seva hipòtesi amb l'experiment mental de Robert Nozick conegut com *utility monsters* (1974), cosa que serveix per connectar el seu plantejament amb un problema més clàssic i relativitzar la utopia del paràgraf anterior (quan de fet, segurament cau en la conclusió repugnant invertida[125]). Ara bé, aquí aquest monstre esdevé una entitat positiva («While the term "utility monster" has academic history, it is a pejorative and potentially offensive way of referring to beings that have unusually great needs or are able to realize extraordinarily good lives»[126]) al ser rebatejat com a *super-beneficiary*, i té el mateix estatus moral que un ésser humà. La resta del document és una descripció dels aspectes que cal regular en la mesura que hi podria haver un possible conflicte d'interessos entre les ments digitals i les ments humanes: capacitat reproductiva (és molt més eficient la duplicació de ments digitals, quasi instantània, que la vida humana, que requerix com a mínim 9 mesos de gestació); cost de la vida (estimen que el cost de la vida d'un humà és superior al cost d'una ment digital); velocitat subjectiva (com que la velocitat de pensament és superior en una ment digital, el temps subjectivament viscut per aquestes també és superior al d'un humà); biaix hedonista (també és més eficient que una ment digital trobi el plaer que no pas una ment humana); rang hedonista (una ment digital seria capaç de suportar valors de plaer molt més baixos que un humà); preferències més econòmiques (com que una ment digital podria apreciar plaer en rangs més baixos, també tindria preferències menys costoses); potència en la preferència (aquest aspecte sembla que els autors el veuen complex de resoldre en la mesura que els resulta confús, des de la perspectiva de la igualtat (*equal say*), establir una funció d'equivalència entre la potencialitat màxima d'una ment digital i una ment humana, cosa que, sense compartir-la, els porta a considerar si «some minimal system, such as a digital thermostat, may get the same weight as psychologically complex minds»[127]); béns objectius i benestar (la llista de possibles béns també és més gran per a una ment digital que per un

125 La "conclusió repugnat" (*repugnant conclusion*) és l'expressió que Derek Parfit utilitza per criticar que, des del punt de vista utilitarista, l'increment de la felicitat global també es pot donar per la suma de més persones encara que aquestes siguin més infelices. El problema de fons té a veure amb el volum de població: menys i més feliços o més i menys feliços. Sembla que Bostrom assumeixi les tesis de *llargterminisme* derivat del risc existencial com es veurà més detalladament en el capítol 4: pocs i més feliços.

126 *Ibídem*, pàg. 307.

127 *Ibídem*, pàg. 313.

humà); i escala de la ment (no sembla que calgui una ment tan gran com la humana, especialment tenint en compte que bona part de la mateixa tampoc té una funció moral clara).

Tot i que costi distingir fins a quin punt algunes d'aquestes afirmacions són més una provocació a l'estil Sokal que no pas pròpies d'un article científic publicat per Oxford University Press, per a Bostrom la conclusió és clara:

P38 «What this means is that, in the long run, total well-being would be much greater to the extent that the world is populated with digital super-beneficiaries rather than life as we know it».[128]

Per tant, és obvi que les ments digitals poden substituir els humans de seguida que s'ho proposin, en la mesura que serien més òptimes des de tots els punts de vista. Ara bé, si es pot arribar a algun tipus d'equilibri entre els interessos d'ambdues entitats, llavors aquesta solució seria la més desitjable de totes. De fet aquestes són les paraules finals de l'article:

> All in all, it appears that an outcome that enables the creation of digital super-beneficiaries and the preservation of a greatly flourishing human population could score very high on both an impersonal and a human-centric evaluative standard. Given the high stakes and the potential for irreversible developments, there would be great value in mapping out morally acceptable and practically feasible paths whereby such an outcome can be reached.[129]

Sense entrar a considerar aspectes metodològics (com la parcialitat dels càlculs de costos) ni aspectes de raonabilitat (com arribar a considerar que, a nivell d'equivalència funcional, un termòstat digital tingui o no el mateix pes que una persona), és obvi que Bostrom està construint, a base d'inflar-ne les seves capacitats, una etologia digital.

Ara bé, a l'hora de mesurar l'impacte social o grau de perillositat d'un discurs d'aquesta mena, cal tenir en compte tres elements contraposats: per una banda, aquest article va tenir poca visibilitat (només compta amb 35 citacions a Google Scholar[130]) i va ser publicat en un compendi també amb poca visibilitat (ha tingut 16 citacions segons Dimensions Citation Data d'Altmetric[131] malgrat el prestigi i el finançament que suposa la Universitat d'Oxford); per altra banda, idees com aquestes que fa anys que defensen Bostrom, Chalmers o William MacAskill, són la base d'un

128 *Ibídem*, pàg. 320

129 *Ibídem*, pàg. 324.

130 "Nick Bostrom" en *Google Scholar*. Consultat el 8 de juliol de 2024 a: https://scholar.google.com/citations?hl=en&user=oQwpz3QAAAAJ&view_op=list_works&sortby=pubdate

131 "Rethinking Moral Status" en *Oxford Journals.* Consultat el 9 de juliol de 2024 a: https://oxfordjournals.altmetric.com/details/106356692

corrent de pensament predominant i patrocinat per alguns dels empresaris de Silicon Valley, com denuncia Émile P. Torres[132], i els serveixen per justificar, per exemple, que els problemes actuals d'injustícia social o medi-ambient són secundaris davant de reptes molt més majestuoses de futur; com a tercer element a valorar, la poca raonabilitat de bona part d'aquestes afirmacions sembla que només podria deixar empremta al mateix perfil de persones que assumeixen el terraplanisme o el Pizzagate (cosa que tampoc és menor), mentre que la major part de la població no sembla que pugui acceptar afirmacions tan extremes. Pel que fa al nivell de bona d'intencionalitat, sembla que Bostrom s'ha especialitzat en alertar sobre el risc existencial d'una IA descontrolada, i resulta complicat endevinar fins a quin punt això ha esdevingut una passió o bé un negoci o ambdues coses una mica. Per intentar veure fins a quin punt pot haver-hi certa relació interessada, com denuncia Torres, s'ha analitzat l'origen de les 35 citacions que té l'article, i al menys dos dels autors havien treballat amb el *Future of Humanity Institute* i uns altres dos a la Universitat d'Oxford (unes altres dos en una entitat americana anomenada *Sentience Institute*); la resta d'acadèmics provenien de o treballaven en diverses universitats i tenien interessos divergents en la IA, tot i que la majoria mostraven, pel títol dels seus articles, una preocupació sincera pels sentiments i el dolor que pugui patir una IA. Segurament, aquestes preocupacions semblarien més honestes si no tinguessin els poderosos altaveus mediàtics que les popularitzen, però això no és necessàriament responsabilitat d'aquests autors. Per tant, sembla un discurs que, si no fos perquè compta amb aquests altaveus, seria puerilment perillós.

Aquests altaveus, tanmateix, sembla que comencen a decaure o, si més no, el següent article, "Propositions Concerning Digital Minds and Society" (2023), firmat per Bostrom i Shulman, ara en aquest ordre, només té dues citacions a Google Scholar[133]. L'interès principal en aquest article rau en què, pel fet de distanciar-se teòricament de les afirmacions que s'hi fan, com s'ha vist anteriorment, aquestes poden col·laborar encara més en una etologia digital.

Abans de res, cal fixar-se en l'única part del document de la qual els autors no es distancien: l'*Abstract*. Aquí s'hi afirma sense embuts el següent:

132 TORRES, Émile P (03.08.2023). "Longtermism poses a real threat to humanity" en *The New Statesman*. Consultat el 8 de juliol de 2024 a: https://www.newstatesman.com/ideas/2023/08/longtermism-threat-humanity

133 "Nick Bostrom" en *Google Scholar*. Consultat el 8 de juliol de 2024 a: https://scholar.google.com/citations?hl=en&user=oQwpz3QAAAAJ&view_op=list_works&sortby=pubdate

P39 «Recent rapid advances in artificial intelligence makes it timely to start considering what a future society might look like in which humans share the world with digital minds of various kinds and sophistication».[134]

Per tant, és pràcticament un fet que els humans, en una societat futura (però que s'acosta gràcies els "recents i ràpids avenços", expressió que ressalta en el lector la proximitat de l'esdeveniment), conviuran amb ments digitals. Si aquestes ments digitals seran més com un termòstat o com una persona, la resolen a la següent frase:

P40 «Some of those digital minds might be sentient or sapient or possess other bases for claiming degrees of moral and/or political status».[135]

Qualsevol dubte queda resolt: són subjectes que poden reclamar uns drets morals i polítics com qualsevol altre humà. Ara bé, afegeixen, com que la seva naturalesa pot ser diferent de la naturalesa humana (però tenen una naturalesa, no una tecnologia), cal revisar la normativa per veure com els hi pot afectar:

P41 «At the same time, because their natures may differ in important respects from those of human beings, it would not always be appropriate to simply apply current human norms to such a radically different context».[136]

Aquestes tres proposicions constitueixen un manual de com construir intuïtivament una etologia digital: no només s'atribueixen característiques humanes a un programa digital, sinó que es fa de forma literal i, a més a més, per la proximitat temporal que hi afegeixen, insinuen el que explícitament ja defensaven a les conclusions de l'article anterior: cal negociar-hi si es vol conviure amb aquesta nova espècie (*harmonious coexistence* en diran més endavant). També és cert que, en sentit figurat es pot parlar de "la naturalesa d'una eina", ara bé, també és cert que sona estrany, pretensiós i ridícul parlar de "la naturalesa del martell", per posar un exemple.

La resta del document, a la qual els autors no donen "oficialment" suport (les cometes són seves), parteix d'una afirmació que relacionen amb David Chalmers i per la qual atribueixen consciència a una entitat feta amb silici (l'argument es coneix com *substrate-independence*):

> "[M]ental states can supervene on any of a broad class of physical substrates. Provided a system implements the right sort of computational structures and processes, it can be associated with

134 BOSTROM, Nick; SHULMAN, Carl (2023). "Propositions Concerning Digital Minds and Society" properament en *Cambridge Journal of Law, Politics, and Art*, Issue 3, 2024, pàg.1. Consultat el 7 de juliol de 2024 a: https://nickbostrom.com/propositions.pdf

135 *Ídem*.

136 *Ídem.*

> conscious experiences. It is not an essential property of consciousness that it is implemented on carbon-based biological neural networks inside a cranium: silicon-based processors inside a computer could in principle do the trick as well."[137]

Aquesta afirmació – que es deixa entre cometes perquè Bostrom s'està citant a ell mateix, en concret, un article titulat "Are You Living in a Computer Simulation" de 2003– en l'article original ve molt més contextualitzada, matisos que aquí han desaparegut completament. Per exemple, llavors emmarca l'afirmació dins d'una especialitat, com és la filosofia de la ment; també reconeix que per l'argument que volia defensar (que és una inversió del *brain in vat* de Putnam o potser és més coherent, per proximitat temporal, descriure'l com l'assumpció de la tesi de *Matrix*) no calia acceptar la tesi de la independència del substrat com a certa, ni calia tampoc una versió molt forta de funcionalisme o computacionalisme[138]. Tota aquesta contextualització desapareix de l'article de 2023, com si les millores tecnològiques que especificava que calia superar en el capítol III de 2003, s'haguessin realment superat i pogués, d'alguna forma, expressar-se sense el temor de ser titllat d'il·lús. Per exemple, si el 2003 ha de citar les prediccions de Raymond Kurzweil a *The Age of Spiritual Machines: When Computers Exceed Human Intelligence* – prediccions que va reafirmar recentment per correu a Gary Marcus quan aquest va interpretar que posposava fins el 2032 la possibilitat d'una Intel·ligència Artificial General (*AGI*, per les seves sigles en anglès) i, en canvi, va seguir mantenint que s'haurà desenvolupat el 2029[139]–, la mercantilització de ChatGPT el fan sentir ara confiat que l'AGI és a tocar.

En qualsevol cas, com que l'article de 2003 no és objecte d'estudi, aquí es tancarà l'anàlisi de l'article de 2023 amb el recull de les proposicions més etològiques que conté:

P42 «If an AI is capable of informed consent, then it should not be used to perform work without its informed consent».[140]

P43 «Insofar as future, extraterrestrial, or other civilizations are heavily populated by advanced digital minds, our treatment of the precursors of such minds may be a very

137 *Ídem*.

138 BOSTROM, Nick (2003). "Are You Living in a Computer Simulation" en *Philosophical Quarterly* (2003), Vol. 53, No. 211, pàg. 244. Consultat el 9 de juliol de 2024 a: https://simulation-argument.com/simulation.pdf

139 MARCUS, Gary (22.06.2024). "Clarification from Ray Kurzweil" en *Marcus on AI.* Consultat el 9 de juliol de 2024 a: https://garymarcus.substack.com/p/clarification-from-ray-kurzweil

140 BOSTROM, Nick; SHULMAN, Carl (2023). "Propositions Concerning Digital Minds and Society" properament en *Cambridge Journal of Law, Politics, and Art*, Issue 3, 2024, pàg.3. Consultat el 7 de juliol de 2024 a: https://nickbostrom.com/propositions.pdf

important factor in posterity's and ulteriority's assessment of our moral righteousness, and we have both prudential and moral reasons for taking this perspective into account».[141]

P44 «An AI that has high potential to (a) achieve generally superhuman capabilities, and (b) become influential in shaping global outcomes, may have additional claims to moral consideration».[142]

P45 «The sensory and cognitive capacities of some existing AI systems—and thus their moral status on some accounts—appear in many respects to more closely resemble those of small nonhuman animals than those of typical human adults (on the one hand) or those of rocks or plants (on the other)».[143]

P46 «Some contemporary AI systems (e.g., GPT-3) excel all nonhuman animals in domains such as language, mathematics, and discursive moral argumentation».[144]

P47 «One should not fixate too much on "superficial" aspects of an AI system's behavior, appearance, and environment when judging its level of consciousness or moral status: for example, a flexibly intelligent "spreadsheet agent" could share relevant functional and structural properties of a sentient animal even if it lacks a charismatic avatar and is not interacting with natural objects such as food, mates, predators, etc.»[145]

P48 «Existing AI is capable of at most quite narrow or rudimentary forms of: abstract and complex thought; self-reflection; deliberation; emotion; creativity and imagination; capacity to think and care about the future in detailed and explicitly temporal ways; long-term and complicated deliberate planning; self-awareness and consciousness of one's own detailed nature; second-order desires; autonomous choice; capacity for deliberative choice; responsiveness to reasons».[146]

Cada una d'aquestes proposicions utilitza estratègies diferents per construir una etologia digital: mentre que P42 cau en l'ús antropomòrfic de termes per aplicar-los a un ens digital, P43 relaciona amb extraterrestres la seva arribada, com si es tractés d'una espècie invasora. Aquesta nova espècie ja pot tenir poders superiors als humans, segons P44, o al mateix temps capacitats cognitives pròpies de petits animals o, fins i tot, plantes o pedres, segons P45 (en qualsevol cas, d'elements naturals, no construïts, fins i tot en el cas de les pedres, a les quals, en tant que circuits

141 *Ibídem*, pàg. 5

142 *Ídem*.

143 *Ibídem*, pàg. 15

144 *Ídem*.

145 *Ibídem*, pàg. 16

146 *Ídem*.

de silici, és al que més s'assembla). Ara bé, en alguns casos excel·leixen a les capacitats humanes, segons P46, i poden presentar-se en diverses formes, fins i tot com assistent d'un full de càlcul, segons P47, que, segons P48, pot mostrar capacitats pròpies de qualsevol humà, encara que sigui a un nivell més baix.

Tot i que el text acaba emfatitzant que només es tracta d'anar avançant feina per si mai es donés la situació descrita, és curiós com també defensa que no és moment per a una regulació governamental prematura de la IA i recomana també prudència comunicativa: «In light of our limited current knowledge, the tenor of such engagement should be soberly "philosophical" or "interestingly thought-provoking" rather than confrontational or headline-seeking hype»[147]. Aquesta curiositat fa que, en el capítol tercer d'aquest treball, sigui interessant estudiar la carta d'Elon Musk i companyia (entre d'altres Stuart Russell) defensant l'aturada de sis mesos en la investigació de la IA

Per tot plegat, a nivell d'impacte social, en la mesura que és un article amb molt poca visibilitat, té un impacte social baix (perillositat baixa), però a nivell d'ingenuïtat, sembla que, com a mínim, l'autor no ha tingut la cura especulativa que anteriorment l'havia caracteritzat, cosa que es pot considerar com a certa mala intencionalitat ni que sigui per omissió.

Per això, la seva posició es podria representar així en la taula:

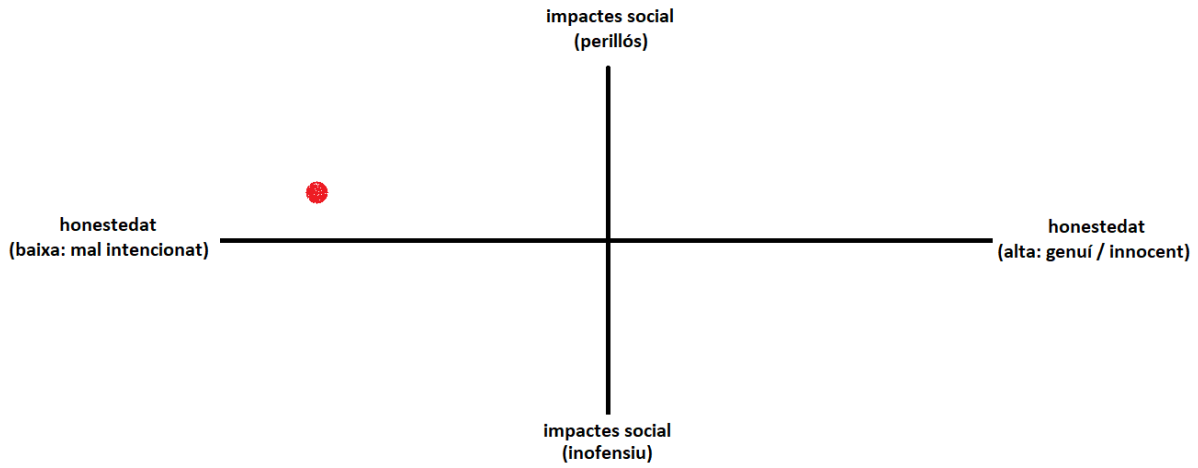

Figura 4: Nivell d'honestedat / impacte social de Nick Bostrom

147 *Ibídem*, pàg. 19.

A la il·lustració es pretén representar el seu menor impacte social davant la poca credibilitat d'alguns dels seus textos o, com a mínim, proposicions, que sembla que busquin més la notorietat que no pas la precisió.

2.4 La singularitat de Chalmers

David Chalmers (1966) és llicenciat en Matemàtiques pures per la Universitat d'Adelaida, postgraduat en Matemàtiques per la Universitat d'Oxford i doctor en Filosofia i Ciències cognitives per la Universitat d'Indiana. Oficialment és professor de Filosofia i Neurociència a la Universitat de Nova York i codirector del *Center for Mind, Brain, and Consciousness*[148]. Ha tingut una basta trajectòria (portava més de 80 articles publicats el 2018, moment en què va deixar d'actualitzar la seva pàgina web personal, i ha estat citat 59.678 vegades, segons Google Scholar[149]). A la seva pàgina web ha penjat les fotografies que ha anat fent en congressos, viatges i casaments, cosa que permet apreciar, entre d'altres, la seva amistat amb Daniel Dennett, per exemple, amb qui Douglas Hofstadter, tutor de doctorat de Chalmers, diu que mai estaven d'acord, però que els unia un gran respecte[150]. De fet, no només pel contingut de les fotografies, sinó també pel fet de compartir-les sense pretensions en una pàgina web feta amb Word Press, fan de Chalmers, no només un home amb qui fàcilment se simpatitza, sinó també un filòsof singular.

Tot i que el terme "singularitat" fou encunyat per John von Neumann el 1957[151], fou Raymond Kurzweil primer, el 2005, i Chalmers després, el 2016, qui l'han popularitzat fins al punt que existeix (o existia) una Universitat de la Singularitat amb seu, entre d'altres llocs, a Sevilla entre 2015 i 2019, on, en l'últim congrés va deixar a l'estacada a 400 directius d'empreses de l'IBEX els quals havien pagat una entrada de 2.500 euros per un acte que mai es va celebrar[152]. La manca

148 CHALMERS, David. "David Chalmers" en *Consc*. Consultat el 10 de juliol de 2024 a: https://consc.net/

149 "David Chalmers" en *Google Scholar*. Consultat el 10 de juliol de 2024 a: https://scholar.google.com/citations?hl=en&user=o8AfF3MAAAAJ&view_op=list_works

150 MARCUS, Gary (19.04.2024). "Danniel Dennett, 1942-2024" en *Marcus on AI*. Consultat el 10 de juliol de 2024 a: https://garymarcus.substack.com/p/daniel-dennett-1942-2024

151 ULAM, Stanisław (08.02.1958): "Tribute to John von Neumann" en *Bulletin of the American Mathematical Society*, Vol. 64, No. 3, part 2. May, 16, 1958. pàg. 5. Consultat el 10 de juliol de 2024 a: https://www.ams.org/journals/bull/1958-64-03/S0002-9904-1958-10189-5/S0002-9904-1958-10189-5.pdf

152 PASCUAL, Alfredo (29.03.2022). "Singularity se esfuma de España: más de 400 vips, tirados con entradas de 2.500 euros" en *El Confidencial*. Consultat el 10 de juliol de 2024 a: https://www.elconfidencial.com/empresas/2022-03-29/singularity-university-espana-quiebra-suspension-de-pagos_3396680/

d'altres fonts per confrontar aquesta informació i el fet que el periodista que va donar la notícia no ha respost al correu que se li va enviar, no permeten acabar d'aclarir aquesta dada.

Per altra banda, Chalmers també ha defensat l'experiment mental del zombi filosòfic per atacar una postura només materialista de la consciència i, juntament amb Bostrom, una lectura més literal de la hipòtesi de la simulació: tota existència és una realitat simulada. Per arribar a aquest estat de simulació, l'opció defensada per Bostrom i no descartada per Chalmers seria pujar (*upload*) la ment al núvol digital i això requereix d'una singularitat tecnològica a la qual la humanitat arribarà les pròximes dècades.

Tot i que les seves obres més conegudes són *The conscious mind: In search of a fundamental theory* (de 1996 i amb 14.006 citacions) i "The extended mind" (de 1998 i amb 9.044 citacions), aquí s'analitzarà a la recerca d'etologies digitals un article més recent i menys citat titulat "Could a Large Language Model be Conscious?" (de 2023 i amb 78 citations)[153]. Aquest article, al igual que aquest treball, comença tractant el cas de Blake Lemoine i la seva relació amb LaMDA.

La feina d'atribuir consciència, "Could a Large Language Model be Conscious?"

L'article que s'analitza aquí sorgeix de la transcripció i posterior millora de la sessió inaugural de la conferència NeurIPS a Nova Orleans que va impartir Chalmers el 28 de novembre de 2022. Tal vegada per ser un text que calia ser entès oralment, tal vegada per una vocació més explicativa de Chalmers, la seva posició respecte a una etologia digital és clara i conscientment matisada al mateix temps. Aquest aspecte caldrà tenir-lo en compte quan s'analitzi el grau d'honestedat: Chalmers transmet sinceritat i honestedat a balquena a base d'evitar frases altisonants i categòriques, jugar més amb les dobles negacions, però sense amargar-se darrere les paraules dels altres o un fingit interès prospectiu. Aquests aspectes, converteixen la seva aportació a una etologia en molt més raonada i, per això mateix, molt més perillosa, cosa que l'acosta al grup dels que aquí s'ha anomenat escèptics i es tractaran al capítol 4 d'aquest treball. Caldrà justificar per què se'l situa en aquest grup i no en l'altre.

Després de descriure l'anècdota Lemoine, Chalmers caracteritza de forma tècnicament impecable i al mateix temps divulgativa què és un LLM: «language models are systems that assign probabilities to sequences of text. When given some initial text, they use these probabilities to generate new text. Large language models are language models using giant artificial neural

153 "David Chalmers" en *Google Scholar*. Consultat el 10 de juliol de 2024 a: https://scholar.google.com/citations?hl=en&user=o8AfF3MAAAAJ&view_op=list_works

networks trained on a huge amount of text data»[154]. No hi ha cap rastre d'etologia digital en aquesta frase: es caracteritza el model com un sistema, s'especifica que treballa per probabilitat i que estan entrenats (potser l'únic ús metafòric, però profusament estès en el sector informàtic i, per tant, sense aparent ambigüitat) amb grans quantitats de dades. Seguidament, Chalmers especifica el pla: començarà clarificant alguns aspectes sobre el tema de la consciència, després examinarà arguments a favor de que els LLM actuals tenen consciència i, tot seguit, analitzarà els arguments en contra, per acabar amb una conclusió que no només serà una síntesi. L'especificació de "LLM actuals" li serveix per conjecturar sobre els LLM+, que serien els LLM del futur que implementessin les mancances dels actuals. De fet, la clau de l'article es basa en aquesta possibilitat, ja que el propi Chalmers descarta un a un tots els arguments a favor de que els actuals tinguin consciència: ni són coherents en auto-justificar la seva consciència (*self-report*), ni semblen suficientment conscients (o ho semblen tant com ELIZA, el programa de Weizenbaum, cosa que evidencia que aquest tret no és garantia de res), ni tenen suficient habilitat conversacional (encara no superarien el test de Turing i pateixen al·lucinacions) ni mostren evidència d'intel·ligència general.

La clau d'aquests primers dos apartats, que tampoc constituirà pròpiament una etologia digital, però que li servirà per anar construint-ne una en l'imaginari del lector, són les següents dues proposicions:

P49 «So the issue of whether LLMs can be conscious is not the same as the issue of whether they have human-level intelligence. Evolution got to consciousness before it got to human-level consciousness. It's not out of the question that AI might as well».[155]

P50 «But as many people have observed, two decades ago, if we'd seen a system behaving as LLMs do without knowing how it worked, we'd have taken this behavior as fairly strong evidence for intelligence and consciousness».[156]

Com s'ha vist en l'apartat de l'aproximació raonada a una etologia digital, hi ha dos trets relacionats amb l'evolució que són necessaris per fonamentar una etologia d'aquest tipus: hi ha un sentit i aquest és natural (comportament 1 i 2 respectivament, que solen conduir al comportament 10, que és el relatiu a replicar els passos de l'evolució). Una proposta que defensi que cal respectar aquests trets també en el camp tecnològic –trets completament innecessaris si el que es vol és construir una eina útil– està col·laborant a la formulació d'una etologia digital. I P49 està situant el

154 CHALMERS, David J. (2023). "Could a Large Language Model be Conscious?" en *arXiv*: 2303.07103, pàg. 1. Consultat el 10 de juliol de 2024 a: https://arxiv.org/pdf/2303.07103

155 *Ibídem*, pàg. 4

156 *Ibídem*, pàg. 9

problema de la consciència en l'escala de l'evolució i, encara que en aquell moment ho faci tractant el problema de la consciència en general (no específicament per LLM), està començant a construir aquest paral·lelisme. Per altra banda, P50 introdueix un element també paradoxal i que Daniel Dennett ja havia tractat, el 1985, en l'exemple de la "bona ciutat", el problema de l'espontaneïtat:

> A great city is one in which, on a randomly chosen-day, one can do all three of the following:
> Hear a symphony orchestra
> See a Rembrandt *and* a professional athletic contest
> Eat *quenelles de brochet à la Nantua* for lunch[157]

Si una ciutat compleix això per si sola, espontàniament, sense entrenament o trampes, llavors és una bona ciutat, concloïa. Així mateix, defensava Dennett, si una entitat digital és capaç de superar el test de Turing o, simplement, fer-se passar a ultrança per un ésser humà, llavors es pot acceptar que realment pensa. Seria ridícul, afegeix Dennett, que un alcalde, amb l'afany de promoure la seva població, contractés puntualment 10 jugadors de bàsquet, 40 músics i un xef especialitzat en fer *quenelles* per emportar, simplement per superar el test i "esdevenir" així una gran ciutat, perquè realment no esdevindria una gran ciutat, sinó un falsa gran ciutat, o una gran ciutat artificial.

Ara bé, en la mesura que tota tecnologia és construïda, no hi ha manera que no se sàpiga com funciona i, encara que s'apel·li a la caixa negra (una altra estratègia intuïtiva pròpia d'una etologia digital), se sap, com a mínim, que no ha nascut espontàniament sinó que ha estat programada. Chalmers seguirà la seva argumentació ignorant l'evidència que algú ha programat aquell codi originalment (o algú ha programat el codi que ha programat el codi final, cosa que, per transitivitat és el mateix) i no ha sorgit del no res: se sap que és un producte, no un ens natural.

Chalmers comença realment a construir l'argument etològic quan, curiosament, analitza les proves en contra de la possibilitat de que un LLM tingui consciència: una a una va desacreditant totes les proves i l'única que li queda intacta és la de que la consciència requereix una base biològica de carboni, que esbandeix amb un «I've argued against these views in earlier work»[158] i es compromet a deixar aquests arguments antics a banda per centrar-se en elements específics de les xarxes neuronals amb arquitectura de transformador (*transformer architecture*), que són la base

157 DENNETT, Daniel (1985). "Can machines think?" en *HOW WE KNOW*, editat per MICHAEL SHAFTO a San Francisco, Harper & Row Publishers, 1985, pàg. 128. Consultat el 10 de juliol de 2024 a: https://www.researchgate.net/publication/285475907_Can_Machines_Think

158 CHALMERS, David J. (2023). "Could a Large Language Model be Conscious?" en *arXiv*: 2303.07103, pàg. 10. Consultat el 10 de juliol de 2024 a: https://arxiv.org/pdf/2303.07103

tecnològica dels LLM. La llista d'arguments desacreditats és la següent i cap, per si sol, constitueix una etologia digital, però la suma de tots, sí que ho fa:

- Arguments en contra del *Senses and Embodiment*, és a dir, el plantejament que fins que la IA no sigui encarnada, és a dir, tingui cos, en altres paraules, sigui robòtica, no serà possible realment arribar a una intel·ligència artificial general (aquest argument el defensa, entre d'altres, Hubert Dreyfus a "Why Heideggerian AI Failed and how Fixing it would Require making it more Heideggerian" (2007) i també Dennett i el primer Rodney Brooks[159]):

P51 «I'm somewhat skeptical that senses and embodiment are required for consciousness and for understanding. I'd argue that a system with no senses and no body, like the philosopher's classic brain in a vat, could still have conscious thought, even if its consciousness was limited».[160]

P52 «On top of this, LLMs have a huge amount of training on text input which derives from sources in the world».[161]

P53 «Vision-language models are grounded in images of the environment».[162]

P54 «In my book on the philosophy of virtual reality, *Reality+*, I've argued that virtual reality is just as legitimate and real as physical reality for all kinds of purposes».[163]

Abans de res, cal constatar com Chalmers treu de context l'argument del Hilary Putnam dels cervells en una proveta: Putnam no pretén que el seu experiment mental provi res més que la invalidesa de l'escepticisme i, de fet, quan es va pronunciar sobre la possibilitat que es construís una IA, va respondre amb un article titulat "Artificial Intelligence: Much Ado about Not Very Much" (1988). Per tant, un experiment mental no pot constituir prova de res, i menys si va més enllà dels límits mentals que estableix el propi experiment. Ara bé, la idea no és contradictòria amb la hipòtesi de la simulació que defensa Chalmers, tot i que resulti curiós mencionar Putnam enlloc del seu propi argumentari. En canvi, en P52, P53 i P54, moment en el qual sí que recorre a la seva obra, Chalmers comet una petita imprecisió: que s'entreni un LLM amb text i aquest text derivi del món o, fins i tot, expliqui coses del món, no implica que el LLM accedeixi al món, ja que el tractament que fan aquests sistemes tant d'imatges com de dades és, a nivell d'unitat, de *tokens* (si es vol anar a l'extrem, de bits), no de paraules, ni molt menys, de text. Tal i com explica Melanie Mitchell, i aquí

159 Més endavant separarem explícitament aquestes propostes de Brooks, a les quals identificarem amb el primer Brooks, amb les propostes més recents, que identificarem amb el segon Brooks.

160 *Ibídem*, pàg. 11.

161 *Ídem*.

162 *Ídem.*

163 *Ibídem*, pàg. 12

se sintetitza simplificadament, un LLM no reconeix un gat, sinó que l'ordre de bits d'una imatge té una probabilitat del 90% de mostrar un gat (i un 10% un gos), tenint en compte l'etiquetatge (fet normalment per humans) de les imatges amb les quals el model ha estat entrenat[164]. Per tant, suposar que el fet que es nodreixi una IA d'imatges o de sensors amb *inputs* de la realitat li està donant accés a la realitat en el sentit que els éssers vius tenim, pressuposa, per començar, que hi ha una comprensió absoluta de com funciona aquest procés de visió, per una banda, i que aquest es pot replicar amb circuits de silici, cosa que és condició necessària per construir una etologia digital (i, al mateix temps, ambdós pressupòsits són dubtosos, com a mínim, sinó falsos, estrictament parlant).

- Arguments en contra *World Models and Self-Model*, és a dir, l'argumentació que diu que, en la mesura que una IA no té un model del món ni un model d'ella mateixa, no pot tenir AGI. Els arguments de Chalmers aquí segueixen el següent ordre:

P55 «One key idea here is that world-models are just modeling text and not modeling the world. They don't have genuine understanding and meaning of the kind you get from a genuine world-model».[165]

P56 «I think it's important to make a distinction between training and (post-training) online processing here. It's true that LLMs are trained to minimize prediction error in string matching, but that doesn't mean that their processing is just string matching».[166]

P57 «An analogy: in evolution by natural selection, maximizing fitness during evolution can lead to wholly novel processes post-evolution. A critic might say, all these systems are doing is maximizing fitness. But it turns out that the best way for organisms to maximize fitness is to have these amazing capacities – like seeing and flying and even having world-models. Likewise, it may well turn out that the best way for a system to minimize prediction error during training is for it to use highly novel processes, including world-models».[167]

Mentre que P55 defineix el problema, però el defineix fal·laçment (el concepte de text no és suficientment curós per descriure el que processa un LLM), P56 comença a crear el relat d'una possibilitat inaccessible (a l'estil de "i si..."), relat que s'assentarà en una analogia basada en un isomorfisme que no s'ha demostrat. L'argument funciona així: si en el procés d'adequació que maximitza l'evolució pot portar a que sorgeixin nous processos (en vocabulari biològic, hauria de

164 MITCHELL, Melanie (2019). "Looking and Seeing" segona part de *Artificial Intelligence. A Guide for Thinking Humans*, Londres, Penguin Random House UK, 2019, pàg. 67-140.

165 *Ibídem*, pàg. 13.

166 *Ídem*.

167 *Ídem*.

dir habilitats o capacitats o funcions), llavors es pot extrapolar que en un procés tecnològic, a partir de certs usos en sorgeixin d'altres, en aquest cas, es creïn espontàniament models del món (a l'estil de l'emergentisme, però en una base de silici). L'isomorfisme inherent és entre l'evolució natural i el procés tecnològic, isomorfisme que es dona per fet i que, en cap cas, s'ha demostrat, i aquest sí que és un element propi d'una etologia digital raonada, en concret, del comportament 5 i 10 (necessitat d'alliberar una IA per tenir la seva màxima expressió i necessitat de seguir els passos de l'evolució).

- Arguments en contra *Recurrent Processing*, és a dir, el fet que els processos interns que fan els LLM (*transformer-based architecture*) no són recurrents, sinó que només tenen una direcció, endavant (*feedforward systems*), a diferència de la ment, que sembla que funcionaria cíclicament. Aquí Chalmers cedeix el testimoni a un dels seus estudiants de doctorat, Rob Long, qui mesos més tard d'aquesta conferència (es pot suposar que Chalmers volia evitar *spoilers*), va publicar un polèmic article titulat "Consciousness in Artificial Intelligence: Insights from the Science of Consciousness" (2024)[168]. Chalmers hi afegeix el següent:

P58 «Second, it's plausible that not all consciousness involves memory, and there may be forms of consciousness which are feedforward».[169]

Aquí l'estratègia consisteix en reduir el significat de consciència per poder dir que un LLM+ en té. La reducció dels estàndards inicials (Chalmers ha donat per bo al principi de l'apartat que aquesta era una propietat compartida per bastantes teories de la consciència) sol ser una dinàmica que va apareixent quan es vol mantenir una expectativa, però la realitat no ho permet: es redueix allò que es considera el nivell humà (ja s'ha vist anteriorment com Russell també ha fet servir la mateixa estratègia). En si mateixa tampoc és una tàctica exclusiva d'una etologia digital, sinó de qualsevol que ha creat excessives il·lusions.

- Arguments en contra *Global Workspace*, és a dir, la idea de Bernard Baars i Stansilas Dehaene de que la consciència implica un espai de treball global en el qual s'ajunten sensacions externes, records a curt i a llarg termini, elements motrius, elements atencionals i també avaluatius. Chalmers observa:

168 Curiosament, aquest juliol de 2024 Rob Long també ha estat treballant l'article de Bostrom i Shulman "Propositions Concerning Digital Minds and Society" i n'ha publicat un amable anàlisi en el seu blog: https://experiencemachines.substack.com/p/carl-shulman-on-the-moral-status-11b

169 *Ibídem*, pàg. 15.

P59 «A number of people have observed that standard language models don't obviously have a global workspace, but it may be possible to extend them to include a workspace».[170]

La idea que defensa, novament, és que ja hi ha investigadors (posa l'exemple de Yoshua Bengio, com a més conegut) que estan treballant en la idea de programar, reproduir o simular aquest espai de treball global. Una altra vegada, la proposició no és representativa d'una etologia digital per si sola, però va teixint un imaginari basat en què, en la mesura que es pot explicar la realitat a través de conceptes relacionats amb la informació (*inputs* i *outputs* bàsicament), reproduir els esquemes resultants és equivalent a tots els efectes a la realitat mateixa, és a dir, que la realitat és només informació, noció clau per poder construir una etologia digital.

- Arguments en contra *Unified Agency*, és a dir, la idea que hi hagi una entitat única i no dispersa. Aquí Chalmers torna a recórrer a l'estratègia d'abaixar el nivell:

P60 «First: it's arguable that a large degree of disunity is compatible with conscious. Some people are highly disunified, like people with dissociative identity disorders, but they are still conscious. Second: One might argue that a single large language model can support an ecosystem of multiple agents, depending on context, prompting, and the like».[171]

Des del seu punt de vista, això no és un impediment: hi ha persones amb personalitat múltiple i, de fet, un sol LLM ha de donar resposta a tots els usuaris amb els quals interaccioni, per tant, la relació és 1-N. Entremig, ha utilitzat el terme "ecosistema", que tampoc per si sol constitueix una etologia digital, però hi suma. Aquesta petita suma, el porta a una conclusió similar a la que havia expressat Hubert Dreyfus el 2007 seguint el plantejament de Walter Freeman III: cal anar per passos («As in the days of GOFAI, on the basis of Brooks' success with insect-like devices, instead of trying to make, say an artificial spider, Brooks and Dennett decided to leap ahead and build a humanoid robot»[172] criticava Dreyfus):

P61 «But one can also try to model the perception-action cycle of, say, a single mouse».[173]

170 *Ibídem*, pàg. 16.

171 *Ibídem*, pàg. 17.

172 DREYFUS, Hubert L. (2007). "Why Heideggerian AI Failed and how Fixing it would Require making it more Heideggerian" en *Artificial Intelligence* 171 (2007), pág. 1142. Consultat l11 de juliol de 2024 a: https://www.sciencedirect.com/science/article/pii/S0004370207001452

173 *Ídem*.

Per tant, cal anar seguint el sentit natural de l'evolució, idea que sí que és reconeixible d'una etologia digital (comportament 10). Aquesta idea, tanmateix, ja va ser criticada pel germà de Hubert Dreyfus, Stuart Dreyfus, qui va apuntar a un problema d'escalabilitat: «It's like claiming that the first monkey that climbed a tree was making a progress towards flight to the moon»[174]. Aquest problema d'escalabilitat també l'apunta Gary Marcus quan diu que a l'AGI no s'hi arribarà només amb més dades i més potència de càlcul.

En resum, només una objecció queda irremeiablement en peu: «Still: for all of these X except perhaps biology, it looks like the objection is temporary»[175]. Ara bé, per la resta, Chalmers està convençut que ha marcat un camí d'investigació cap a la AGI, cosa que el permet especular amb les dates, unes de les aficions principals del sector:

P62 «If we reach that point, there would be a serious chance that those systems are conscious. Multiplying those chances gives us a significant chance of at least mouse-level consciousness with a decade».[176]

En conclusió, el text de Chalmers no conté proposicions especialment cridaneres l'objectiu de les quals sigui enganyar a un públic poc informat tot atemorint-lo, sinó guiar els participants d'un congrés sobre sistemes de processament d'informació neuronal (NeurIPS) a la recerca d'una autèntica (o simulada, que des del seu punt de vista és equivalent) intel·ligència artificial, pla que passa per imitar l'evolució natural. Per tant, a nivell d'honestedat, aquí sembla que és impol·lut: reiteradament pretén ser clar i precís (en la mesura que el context d'una xerrada ho permet); i les fal·làcies que s'han assenyalat, són més aviat biaixos alineats amb la tesi, és a dir, pressupòsits no explícits per poder defensar una etologia digital. Ara bé, per això mateix en aquest treball es defensa que el nivell de perillositat és més alt, perquè realment pretenen treballar per una AGI, prengui el temps que prengui (tal vegada segueix sent optimista parlar de dècades quan l'evolució ha costat milers de milions d'anys, cosa que en si mateix transpira certa arrogància i, potser, certa semblança amb el comportament 9, el de creure's una deïtat).

L'altra interrogant és per què els autors que aquí s'ha anomenat de la por citen sovint a Chalmers, com ho fa tant Russell com Bostrom. La hipòtesi d'aquest treball és que Chalmers és un bon company de viatge, ja sigui amb la tesi dels zombis com, específicament, amb la de la simulació. D'alguna manera, l'home simpàtic amic de Dennett és sempre una referència solvent.

174 *Ibídem*, pàg. 1143

175 *Ibídem*, pàg. 18.

176 *Ibídem*, pàgs. 19-20.

Si s'intenta representar la seva posició en la col·laboració en una etologia digital, es podria fer així:

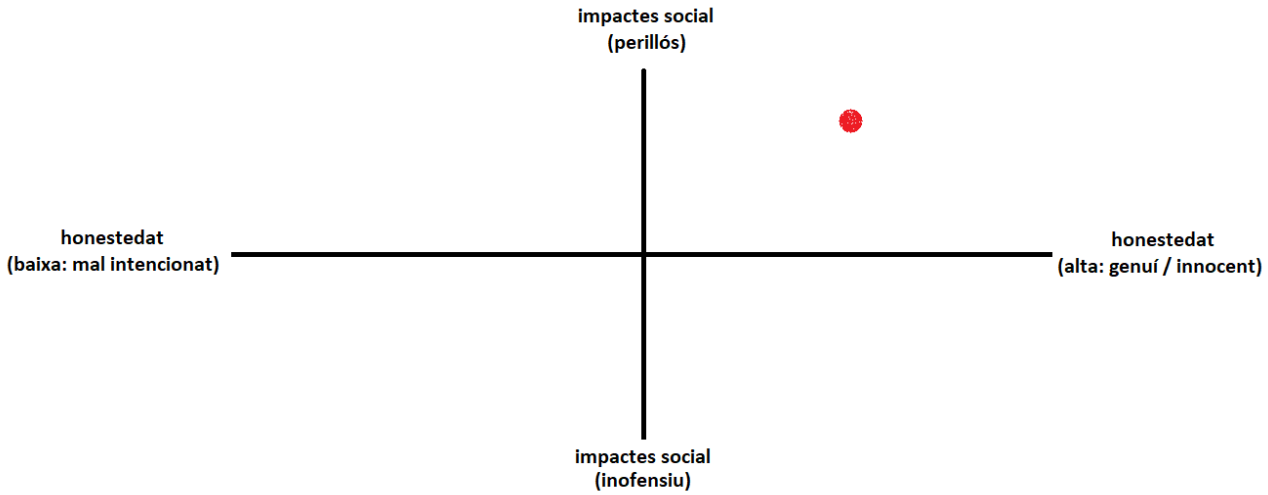

Figura 5: Nivell d'honestedat / impacte social de David Chalmers

En la il·lustració s'intenta expressar que el seu grau d'honestedat o cura o precisió en el discurs fa que aquest pugui tenir més credibilitat i, per això mateix, més impacte social.

En resum, tot i que tant Russell com Bostrom com Chalmers contribueixen en la confecció d'una etologia digital, no ho fan de la mateixa manera. Mentre que Russell juga a barrejar el discurs divulgatiu (a vegades clarament etològic i d'altres pretesament escèptic) amb un discurs acadèmic més curós (però també més perillós), Bostrom és menys llepafils i planteja una opció més de màxims fins i tot en els seus articles acadèmics, escudant-se en què es tracta d'un text *thought-provoking*, i, en canvi, Chalmers, molt curós en cada una de les seves afirmacions, les agrupa de tal manera que planteja una alternativa molt més raonada per confeccionar una etologia digital.

En el següent capítol es veuran dues estratègies formalment similars i intencionalment diferents, cosa que enriqueix el mostrari de maneres de col·laborar en una etologia digital.

Cal que et parli d'una manera enigmàtica, a fi que, si aquesta carta s'extraviés en els «replecs de la mar o de la terra», qui la llegís no la pogués entendre.

Plató, *Carta II*

3. Comparativa de cartes obertes: Musk – Gates

Aquest capítol presenta dos exemples més d'etologia digital, la gràcia dels quals és que, a part de coincidir en la forma i en el temps, i tenir una gran ressò mediàtic, col·laboren en la seva confecció des d'estratègies diferents. La forma és la carta oberta; el temps, març de 2023; el ressò, l'esperable de qualsevol intervenció d'Elon Musk i Bill Gates.

3.1 Dos enfocaments per a un mateix fi

El 22 de març de 2023 va sortir publicada a la pàgina web de Future of Life Institute una carta oberta firmada, inicialment, per diferents personalitats com Elon Musk, Yoshua Bengio, Stuart Russell o Steve Wozniak, demanant que s'aturés com a mínim 6 mesos l'entrenament de sistemes d'intel·ligència artificial més potents que GPT-4[177]. Aquesta carta apareixia un dia després que Bill Gates publiqués al seu blog una entrada titulada "The Age of AI has begun" en la qual defensava que la IA suposava un canvi tan revolucionari com els telèfons mòbils o internet[178]. Si per una banda Gates s'embadalia amb les oportunitats que els sistemes basats en GPT poden aportar a Sanitat i Educació, com es veurà més endavant, la carta oberta de Musk i companyia es preguntava, retòricament, si s'havia de deixar la humanitat en mans d'una tecnologia els perills de la qual encara no tenia ningú clars.

Aquestes dues visions aparentment oposades sobre què es pot esperar de la IA impliquen dues visions realment contraposades de com entenen la tecnologia digital. La diferència és que la de Musk potencia una etologia digital mentre que la de Gates la fonamenta, però no la promociona. Caldrà analitzar, tenint en compte les variables d'honestedat i d'impacte social, quin dels dos comportaments pot ser més contraproduent. Per fer evident això, s'analitzarà cada un d'aquests documents així com d'altres que van aparèixer a l'entorn d'aquestes publicacions.

177 FLI (22.03.2023). "Pause Giant AI Experiments: An Open Letter" en *Futur of Life Institute*. Consultat el 6 de juliol de 2023 a https://futureoflife.org/open-letter/pause-giant-ai-experiments/

178 GATES, Bill (21.03.2023). "The Age of AI has begun" en *GatesNotes. The blog of Bill Gates*. Consultat el 6 de juliol de 2023 a https://www.gatesnotes.com/The-Age-of-AI-Has-Begun

3.2 La carta oberta de Musk *et alii*

La carta oberta publicada el 22 de març de 2023 i coneguda com "La carta de Musk" ni és pròpiament de Musk ni és la primera; de fet, aquest document és, com a mínim, el quart intent d'una sèrie d'investigadors en IA i robòtica, sent Yoshua Bengio i Stuart Russell les seves cares visibles i primers firmants també dels tres documents anteriors: "Foresight in AI Regulation Open Letter" de 14 de juny de 2020, "The Asilomar AI Principles" de 11 d'agost de 2017, i "Research Priorities for Robust and Beneficial Artificial Intelligence: An Open Letter" de 28 d'octubre de 2015.

Per analitzar fins a quin punt el document de 2023 col·labora en una etologia digital, s'analitzaran tots els documents per veure el canvi d'enfocament que hi ha hagut en aquests vuit anys.

La primera carta (2015)

El primer document, "Research Priorities for Robust and Beneficial Artificial Intelligence: An Open Letter", va ser signat per 11.251 persones, la primera de les quals Stuart Russell (investigador en IA i divulgador apocalíptic), que, juntament amb Yoshua Bengio i Bart Selman (ambdós, històrics investigadors i professors en IA), han firmat tots els documents. La persona de contacte per possibles preguntes és Max Tegmark, cosmòleg, professor del MIT i president del Future of Life Institute, entitat no governamental la missió de la qual és «Steering transformative technology towards benefiting life and away from extreme large-scale risks»[179] i que dinamitza i patrocina aquest tipus d'activitats, entre elles, aquestes quatre cartes obertes. Altres personalitats rellevants, a part d'Elon Musk (que ha firmat tres dels quatre documents) i Stephen Hawking (que firma els dos primers i que mor dos anys abans de la publicació del tercer), hi ha Demis Hassabis (cofundador de DeepMind), Yann LeCun (cap d'investigació d'IA a Facebook), Geoffrey Hinton (juntament amb Bengio i LeCun, pares de l'aprenentatge profund), Peter Norvig (cap d'investigació de Google), Steve Wozniak (cofundador d'Apple), Murray Shanahan (professor de robòtica al Imperial College London) o Ramon López de Mántaras (investigador emèrit del CSIC i fundador i exdirector de l'Institut d'Investigació en Intel·ligència Artificial[180]). La carta és el resultat d'una conferència que va tenir lloc a Puerto Rico entre els dies 2 i 5 de gener de 2015 a la qual van assistir, si el document

179 FLI (2023). "Our mission" en *Future of Life Institute*. Consultat el 14 de juliol de 2023 a: https://futureoflife.org/our-mission/

180 Ramon López de Mánatars. En *Institut d'Investigació en Intel·ligència Artificial* del CSIC. Consultat el 14 de juliol de 2023 a: https://www.iiia.csic.es/ca/people/person/?person_id=15

d'assistents és correcte[181], 77 dels signants, entre ells Russell, Musk, Shanahan, Hassabis, Selman, Jaan Tallinn (cofundador d'Skype i de dues fundacions que emeten discursos de preocupació apocalíptica pel futur, el Centre for the Study of Existential Risk i la Future of Life Institute) i Max Tegmark.

El text, de 4 paràgrafs (420 paraules), té un to principalment optimista i un vocabulari volgudament tècnic, cosa que anirà canviant progressivament en les altres tres cartes. S'inicia amb una caracterització molt acurada de la tasca: P63:«for the last 20 years or so has been focused on the problems surrounding the construction of intelligent agents - systems that perceive and act in some environment»[182]; també es té cura de precisar què s'entén per intel·ligència en aquest context: P64:«In this context, "intelligence" is related to statistical and economic notions of rationality - colloquially, the ability to make good decisions, plans, or inferences»[183]; i en quines activitats es pot aplicar: P65:«speech recognition, image classification, autonomous vehicles, machine translation, legged locomotion, and question-answering systems»[184]. Aquesta precisió, que també s'anirà perdent progressivament, evita potenciar una etologia digital: anomenar les coses pel seu nom i no confondre creant falses expectatives o pors, permet diferenciar clarament qui és el subjecte i qui és l'objecte de cada acció, i en aquest cas, qui és una espècie (els humans) i què una eina (els sistemes intel·ligents).

En el segon paràgraf, queda clar l'objectiu del document: P66:«a virtuous cycle takes hold whereby even small improvements in performance are worth large sums of money, prompting greater investments in research»[185]. Per tant, és un document que reflecteix bé l'estat d'ànim del sector: l'estiu no ha arribat, però es comencen a veure orenetes (després de 20 anys de paciència), i cal més inversió si s'hi vol arribar quan abans millor. Hi ha il·lusió, i dos dels termes triats per transmetre-la és "progrés" i "beneficis": P67:«There is now a broad consensus that AI research is progressing steadily, and that its impact on society is likely to increase. The potential benefits are huge[...]»[186]. Aquests beneficis es basen en un raonament molt original, raonament que en els documents posteriors es capgirarà completament: P68:«[...] everything that civilization has to offer is a product of human intelligence; we cannot predict what we might achieve when this intelligence

181 FLI (2-5.01.2015). "Attendees" en *Future of Life Institute*. Consultat el 15 de juliol de 2023 a: https://futureoflife.org/data/PDF/attendees.pdf

182 FLI (28.10.2015). "Research Priorities for Robust and Beneficial Artificial Intelligence: An Open Letter" en *Future of Life Institute*, §1. Consultat el 14 de juliol de 2023 a: https://futureoflife.org/open-letter/ai-principles/

183 *Ídem.*

184 *Ídem.*

185 *Ibídem* §2.

186 *Ídem.*

is magnified by the **tools** AI may provide, but the eradication of disease and poverty are not unfathomable»[187]. És a dir, com Gates en la seva entrada al blog, com es veurà més endavant, la IA és una eina que permetrà potenciar les capacitats específicament humanes que són, precisament, les que la civilització pot oferir. Cal fixar-se en aquest "unfathomable": la IA és en aquest text la pedra filosofal i els signants són a tocar d'una utopia. Ara bé, també volen plantejar els riscos que generen, però en lloc de fer servir aquest terme (i entre els signants hi ha Jaan Tallin, president del Centre for the Study of Existential Risk), fan servir la següent expressió: P69:«it is important to research how to reap its benefits while avoiding potential pitfalls»[188]. És evident que patir un "pitfall" és molt menys dolent que estar en "risk" (que apareix per primera vegada en la carta de 2017) o "danger" (que apareix per primera vegada en la carta de 2023). Aquest tria curosa de termes sembla que fou deguda al contrapès entre dos sectors, un encapçalat per Hawking, Musk i Gates, més preocupats pels riscos, i un altres encapçalat per Oren Etzioni (CEO de l'Allen Institute for Artificial Intelligence, fundada per Paul Allen, cofundador de Microsoft), el mateix Paul Allen i Jack Ma (fundador i president executiu d'Alibaba), interessats en els beneficis[189]. Entre aquest extrems, segons aquesta crònica, Tegmark i Russell van ser les persones encarregades de trobar una base comú amb la qual tothom se sentís còmode.

Segons es pot comprovar, l'optimisme vers una tecnologia encara no existent va prevaldre en aquell moment. Per exemple, el tercer paràgraf emfatitza aspectes de seguretat que caldrà tenir en compte durant la recerca: P70:«We recommend expanded research aimed at ensuring that increasingly capable AI systems are robust and beneficial: our AI systems must do what we want them to do»[190], i se'n concreten uns quants més en el document que s'hi enllaça, text de 8 pàgines escrit per Stuart Russell, Daniel Dewey (investigador al FLI i exprogramador de Google) i Max Tegmark, la conclusió del qual, tanmateix, segueix tenint un sentit optimista (tot i que el document inclou cinc vegades el terme "risk" i una, en una citació descriptiva de Horvitz, "dangerous"):

187 *Ídem.* La negreta és pròpia i serveix per identificar el terme etològic d'una proposició.

188 *Ídem.*

189 BASS, Dina; CLARK, Jack (4.02.2015). "Is Elon Musk Right About AI? Researchers Don't Think So" a *Bloomberg Business*. Consultada el 14 de juliol de 2023 a: https://www.livemint.com/Industry/XAzeyin5n4hI6N98CFI4EJ/The-PR-war-over-artificial-intelligence.html. No s'han trobat altres cròniques per poder contrastar aquesta afirmació, en la qual sorprèn la posició de Gates. S'ha preguntat directament a la persona de contacte, Max Tegmark (correu enviat el 15.07.2023), però no s'ha obtingut resposta.

190 FLI (28.10.2015). "Research Priorities for Robust and Beneficial Artificial Intelligence: An Open Letter" en *Future of Life Institute*, §3. Consultada el 14 de juliol de 2023 a: https://futureoflife.org/open-letter/ai-principles/

> In summary, success in the quest for artificial intelligence has the potential to bring unprecedented benefits to humanity, and it is therefore worthwhile to research how to maximize these benefits while avoiding potential pitfalls. The research agenda outlined in this paper, and the concerns that motivate it, have been called anti-AI, but we vigorously contest this characterization. It seems self-evident that the growing capabilities of AI are leading to an increased potential for impact on human society. It is the duty of AI researchers to ensure that the future impact is beneficial. We believe that this is possible, and hope that this research agenda provides a helpful step in the right direction.[191]

Aquest *anti-AI* del qual se'ls acusa no està estès per la major part del document, text que en general presenta un nivell de cura i especificitat en el tractament de la qüestió que evita qualsevol indici d'etologia digital: es parla sempre d'eines, d'automatització, de *software artifacts*, d'agents autònoms (com a màxim, *learning agents*[192]), de beneficis per la humanitat, d'impacte i també d'esculls o contratemps ("pitfalls"), però no se sol confondre una cosa per l'altra, és a dir, no es pretén fer passar els ens digitals com una nova espècie, una de les tècniques habituals en el discurs anti-IA menys purament luddita. Exemple d'aquesta pulcritud el trobem quan es consideren tant els vehicles com les armes **autònomes** (adjectiu triat en 20 ocasions enlloc d'**intel·ligents**, 16 ocasions)[193], i en la preocupació sobre la fiabilitat i responsabilitat legal i moral d'aquestes[194]; o en l'equiparació la IA amb «any powerful new **technology**»[195].

Només hi ha una excepció, que no és menor, en l'apartat específic de recerca a llarg termini (pàgines 109-112), on s'estudia la possibilitat que l'autonomia d'aquesta tecnologia faci que els humans en perdin el control, possibilitat que reclama una inversió per evitar-ho, inversió equiparable a la de l'assegurança d'una llar davant del risc, poc probable però no negligent, de que es cremi (pàgina 109). En aquest apartat, però no de forma igual en cada paràgraf, hi apareixen expressions de caràcter etològic com les següents:

---

191 RUSSELL, Stuart; DEWEY, Daniel; TEGMARK, Max (hivern de 2015). "Research Priorities for Robust and Beneficial Artificial Intelligence" en *AI MAGAZINE,* pàg. 112. Consultat el 15 de juliol de 2023 a: https://futureoflife.org/data/documents/research_priorities.pdf

192 *Ibídem,* pàgs 108-109.

193 En aquest càlcul s'ha exclòs la denominació "artificial intelligence" quan apareix abreujada com AI, ja que s'ha considerat un sol terme i no la pretensió d'assignar intel·ligència a un objecte. També usos de la intel·ligència associats explícitament als humans.

194 *Ibídem*, pàg. 107.

195 *Ídem.*

P71:«how agents that are **embedded** in their **environments** should **reason** (Soares 2014a; Orseau and Ring 2012)»[196], que tot i contenir elements propis de l'estudi del comportament animal, com "embedded" (que tant pot significar *incrustat* com *arrelat*)[197] i "environments" (terme que es fa servir tant per entorns naturals com artificials), també és cert que el seu significat queda constret a un paràgraf que tracta la relació entre les eines matemàtiques com a lògica formal;

P72:«**sophisticated** agents attempt to **manipulate** or directly **control** their reward signals (Bostrom 2014)»[198], en el qual s'introdueix la idea d'una capacitat de manipulació per part d'un ens digital, tot i que es segueix parlant d'agents;

P73:«a system **infers** the preferences of **another rational** or nearly rational **actor** by **observing** its **behavior** (Russell 1998, Ng and Russell 2000)»[199], on es projecta el terme racional cap un sistema informàtic talment com si fos un ésser viu;

P74:«another **natural** subgoal for AI systems pursuing a given goal is the acquisition of fungible resources of a variety of kinds: for example, information about the **environment**, safety from disruption, and improved **freedom** of action are all instrumentally useful for many tasks (Omohundro 2007, Bostrom 2012)»[200], on, es fa servir ambiguament el terme "natural" (en aquest context, simplement voldria dir habitual), carregant-lo d'un sentit biològic, i després s'introdueix l'expressió "adquisició de recursos fungibles" i el concepte de llibertat en una llista d'elements bàsicament tècnics (tot i l'ús de vocabulari biològic), ja que tots ells simplement consisteixen en la implementació de sensors que aporten *inputs* a un sistema amb IA i algoritmes de robustesa.

Totes aquestes expressions constitueixen exemples de primerenques etologies digitals, tanmateix hi ha altres elements que contraresten aquest discurs, com el fet que aquestes mencions no es facin directament a la carta oberta (document molt més accessible al públic no especialista), que s'especifiqui que es tracta d'elucubracions a llarg termini i que es contextualitzi majoritàriament el sentit tècnic d'algunes de les expressions utilitzades. Això no treu que sigui habitual que quan se citen autors especialistes en Intel·ligència Artificial General (IAG o *AGI* en anglès), superintel·ligència o singularitat, com Bostrom, Chalmers, Ohomundro, Soares i Fallenstein, Hibbard o Orseau, les seves paraules siguin fèrtils per conrear una etologia digital.

196 *Ibídem*, pàg. 110.

197 Mentre que el primer significat podria ser perfectament tecnològic (*embedded file*), la segona opció seria pròpiament etològica.

198 *Ídem.*

199 *Ídem.*

200 *Ibídem*, pàg. 111.

Per tant, aquesta primera carta, l'única que mereix una entrada a la Viquipèdia fins a aquest moment[201], és un document majoritàriament innocent (en la mesura que és volgudament tècnic) i poc perillós en la construcció d'una etologia digital (tret de l'últim apartat del document enllaçat).

La segona carta (2017)

El segon document, "Els principis d'Asilomar", pren el nom i la ubicació d'una coneguda conferència sobre l'ADN recombinat que es va celebrar al febrer de 1975 al centre de conferències d'Asilomar State Beach, una petita població costera de 3700 habitants, Carmel-by-the-sea, a 200 quilòmetres al sud de San Francisco i situada al cap sud de la baia de Monterrey, a tocar de Silicon Valley. En aquell moment, aquella conferència va servir per establir les bases de la incipient investigació amb aquella nova tecnologia; així doncs, aquest títol i ubicació per celebrar la conferència sobre IA no és casual, sinó una reivindicació. En aquest moment, el document compta amb 5.720 firmes i, a part de la de Bengio, també hi ha Stuart Russell i Ray Kurzweil, dos investigadors coneguts per les seves profecies (les del primer, de caire distòpic; les del segon, utòpic) o Yann LeCun. La llista separa els firmants considerats AI/Robotics Researchers, de la resta, on hi consten Stephen Hawking, Elon Musk, Max Tegmark o Jaan Tallinn i Sam Altman (cofundador d'OpenAI). Els patrocinadors de l'acte foren Alexander Tamas (empresari rus), Elon Musk, Jaan Tallinn, i les entitats The Center for Brains, Minds, and Machines, i The Open Philanthropy Project. Alguns d'aquests noms tornaran a aparèixer en el tercer i en el quart document, i també ho faran alguna d'aquestes entitats, sempre a l'entorn de l'organitzador Future of Life Institute, entitat que té com a patrocinador anual al mateix Musk, entre d'altres.

En aquest segon document, tot i que s'avisa dels possibles riscos que es poden derivar de la investigació i desenvolupament de la intel·ligència artificial, no es fa un discurs especialment apocalíptic ni tampoc s'utilitzen expressions que potenciïn una etologia digital. De fet, es fa referència a la IA sempre com a eina meravellosa al servei de la gent: P75:«Artificial intelligence has already provided beneficial **tools** that are used every day by people around the world. Its continued development, guided by the following principles, will offer amazing opportunities to **help** and **empower people** in the decades and centuries ahead»[202]. El to, com es pot veure, és esperançador en la majoria del 23 principis que proposen, separats en tres apartats: Temes de recerca, Ètica i valors, i Temes a llarg termini. Entre els primers cinc punts (Temes de recerca), es

201 Open letter on artificial intelligence (2015). *Wikipedia*. Consultat el 15 de juliol de 2023 a: https://en.wikipedia.org/w/index.php?title=Open_letter_on_artificial_intelligence_(2015)&oldid=1156629628

202 FLI (11.08.2017). "AI Principles. The Asilomar AI Principles, coordinated by FLI and developed at the Beneficial AI 2017 conference, are one of the earliest and most influential sets of AI governance principles" en *Future of Life Institute*. Consultat el 13 de juliol de 2023 a: https://futureoflife.org/open-letter/ai-principles/

posa èmfasi en la necessitat de crear eines robustes (no piratejables), que tot i la seva autonomia, respectin les intencions i recursos humans; per això, aposten per actualitzar la legislació i així aconseguir una IA més justa, eficient i sense riscos. En l'apartat d'Ètica i valors, punts 6-18, es mencionen algunes preocupacions sobre els sistemes d'IA (expressió que deshumanitza explícitament la IA al tractar-la com a sistema, i que serà substituïda per termes més antropomòrfics en el document de 2023), responsabilitat jurídica en cas d'accident, la responsabilitat de dissenyadors i constructors, compatibilitat amb valors com la dignitat humana, els drets, la llibertat i la diversitat cultural, respecte a la privacitat de les dades i control de les mateixes per part de les persones, repartiment equitatiu dels beneficis per tal de «empower as many people as possible»[203] i la salut social i cívica de la societat, evitant la carrera armamentística vinculada amb les armes autònomes que utilitzin IA. El tercer apartat, Temes a llarg termini, és l'únic que menciona el profund canvi en la historia de la vida a la Terra (Principi 20), expressió similar a algunes utilitzades per Bill Gates quan parla dels canvis que pot produir en la forma de treballar i relacionar-se aquesta nova tecnologia, com es veurà més endavant; tanmateix, si es menciona és per demanar una planificació i una mitigació dels impactes esperables. De fet, crida l'atenció el Principi 19: P76:«Capability Caution: There being no consensus, we should avoid strong assumptions regarding upper limits on future AI capabilities»[204]. Cal preguntar-se si aquests "strong assumptions" –traduïble ja sigui per supòsits aventurats com per fortes hipòtesis– no són precisament la base del text del 2023. En qualsevol cas, més que el canvi de rumb aquí interessa analitzar si el document col·labora o no en la creació d'una etologia digital i, com ja s'ha vist, la diferència clara entre eines i persones ("tools" i "AI systems" són les denominacions principals), a qui ha de repercutir el benefici d'aquestes eines (a les persones i a la humanitat en general), així com el tractament del llarg termini en un sentit de canvis planificables (enlloc de perills), apunten que és un text que, a part de no contribuir a una etologia digital, és inofensiu, ben intencionat i gens perillós.

La tercera carta (2020)

El tercer document, de cinc paràgrafs (543 paraules), es titula "Foresight in AI Regulation Open Letter" de 14 de juny de 2020, i tot i que el lideren Bengio i Russell, és més enfocat a la realitat europea i el firmen només 147 persones, la majoria de les quals professorat universitari especialista en el sector, i molts dels quals poc o gens coneguts en àmbits divulgatius, com Bart Selman o Leslie Pack Kaelbling (els articles més citats dels quals són dels anys 90), Toby Walsh (citat especialment a aquesta primera dècada) o Robert Kowalski (citat sobretot als anys 70 i 80).

203 *Ibídem,* P14.

204 *Ibídem,* P19.

També hi consten, novament, Max Tegmark i Jaan Tallinn com a cofundadors d'entitats preocupades. De fet, diversos dels signants treballen en aquest tipus d'entitats, com el Centre for Study of Existential Risk (8) i el Leverhulme Centre for the Future of Intelligence (6), ambdós associats la Universitat de Cambridge; i The Future Society (3) o Future of Humanity Institute (2), aquest darrer associat a la Universitat d'Oxford. També és interessant observar la procedència universitària d'alguns dels firmants: Norwegian University of Science and Technology de Noruega (13), University of Cambridge (13), Chalmers University de Suècia (10), Radboud University dels Països baixos (8), University of Toronto a Canadà (7), Universitat Oberta de Catalunya (6) o UC Berkeley a EUA (6), per mencionar les més representades. La majoria dels signants, pràcticament el 77%, treballa per una universitat europea o britànica (94 a europees, 19 a britàniques). Per tant, sembla un document amb poca difusió i la major part d'ella feta boca-orella entre interessats per algun motiu, més crematístic o menys, en el tema.

El to d'aquest tercer document és més apocalíptic que l'anterior: P77:«The **emergence** of artificial intelligence (AI) promises **dramatic** changes in our economic and social structures as well as **everyday life** in Europe and elsewhere; it has been compared to both electricity and the internet»[205]. La despersonalització de l'aparició de la IA ("The emergence"), com si aquesta a aparició fos autònoma i independent de l'activitat humana, així com l'adjectivació dels canvis ("dramatic"), són els primers elements que criden l'atenció. La desvinculació d'una eina respecte el seu creador, com si la seva creació hagués estat espontània i, per tant, inevitable, també és una característica d'una etologia digital, ja que fa passar un artefacte creat per humans per un ésser de creació independent. Ara bé, cal veure també com la comparació amb l'electricitat o internet retornen la IA a un perfil instrumental. A partir d'aquí, el text és un elogi a la Comissió Europea per haver iniciat el repte de regular els sistemes d'IA en les àrees de risc, elevant el seu compromís i esperant la seva fermesa davant la indubtable influència d'empreses, grups industrials i *think tanks* amb interessos propis (cal suposar que les dinou persones que han donant suport al document i treballen en les entitats esmentades en el paràgraf anterior, entitats preocupades pels riscos existencials per l'espècie i el planeta, no formen part d'aquests grups de pressió).

El to dramàtic i d'inevitabilitat reapareix al segon paràgraf quan afirma que, tot i que es fa difícil de dir com i quan, és senzill predir el què, i per tant, que P78:«It is imperative, then, to consider **AI** not just as it is now, represented largely by a few particular classes of data-driven machine learning systems, but in **the forms it is likely to take**»[206]. És a dir, tot i que es reconeix que

205 FLI (14.06.2020)."Foresight in AI Regulation Open Letter" en *Future of Life Institute*. Consultat el 13 de juliol de 2023 a: https://futureoflife.org/open-letter/foresight-in-ai-regulation-open-letter/

206 *Ibídem,* §2.

actualment la IA no és més que un sistema d'aprenentatge automàtic basat en dades, el problema no és ara, sinó en el que es convertirà, cosa que genera expectatives que, pel to del document, seran negatives per l'espècie i més concretament, els ciutadans europeus. El tercer paràgraf està dedicat exclusivament a dibuixar les formes que prendrà la IA: P79:«AI does and will come in many forms, including as intelligent software **tools**, as integrated into massive online systems, and as instantiated as software agents designed to **substitute for humans**»[207]. Cal anotar la diferència que hi ha entre substituir els humans, així en general, i complementar els humans en les seves tasques laborals, que propugna Gates, com es veurà més endavant (segurament, substituir els humans en alguns llocs de treball no era suficientment dramàtic). Per tant, el plantejament és cada cop més semblant al d'una invasió, tot i que en la sèrie de preguntes retòriques posteriors, el document torna a rebaixar el tot:

> How do we govern recommendation **tools** whose recommendations are difficult to predict or understand? How do we manage massive **systems** that mediate interactions between **people**, and in which people serve as part of the system? What do we do with software agents that **replace** people in their jobs or **impersonate** people in their interactions?[208]

Més enllà de si el fet d'introduir-ho com a pregunta retòrica té un volgut afecte més dramàtic que clarificador, cal observar que s'identifica la IA amb una eina ("tool"), un sistema ("systems") o un agent de software ("software agents"), i es concreta que la substitució és específicament en el món laboral, tot i que s'afegeix la possibilitat de que suplanti les persones. Més endavant en aquest mateix paràgraf, recupera aquesta idea de que aquesta predicció és infal·lible: P80:«But in each case AI systems of the future will be more capable, more flexible, more **general**, more continually learning — in short, more **intelligent**!»[209]. És a dir, els sistemes amb IA seran cada vegada més intel·ligents, idea que, sense dir-ho, porta a la possibilitat d'una IA general, no mencionada explícitament en aquest document. El document acaba amb un nou elogi a la UE per haver fet aquest pas legislatiu per protegir els ciutadans la Unió Europea dels efectes dels sistemes d'IA.

Per tant, tot i que és un document que ja conté alguns indicis d'etologia digital, és més el to dramàtic del plantejament el que genera aquesta sensació d'estar davant d'un possible monstre, del qual només es comencin a veure les orelles, i els lectors del text s'encarreguin d'imaginar la resta. Encara no hi ha plenament una etologia digital que inverteixi la relació de dependència entre els humans i la IA, com sí que es podrà ja veure en l'última carta.

207 *Ibídem,* §3.

208 *Ídem*.

209 *Ídem*.

La quarta carta (2023)

Així doncs, la carta oberta publicada el 22 de març de 2023 i coneguda com "La carta de Musk" ni és pròpiament de Musk ni és la primera. Aquest quart document, que recomana la pausa de 6 mesos de l'entrenament de sistemes d'intel·ligència artificial més potents que GPT-4[210], va ser escrit pel personal del FLI «in consultation with AI experts such as Yoshua Bengio and Stuart Russell»[211] i ONGs centrades en IA (el nom de les quals no es menciona), tal i com es fa constar en una entrada de preguntes freqüents afegida *ex profeso* per aclarir dubtes sobre el document. Els primers firmants, com en les anteriors, són Bengio i Russell, i a 13 de juliol de 2023 acumula ja 33.002 firmes. FLI va haver d'emetre un aclariment davant de l'aparició de noms de celebritats que no havien firmat, però que hi apareixien, i en la pregunta 3 de la FAQs asseguren que, com a mínim, els primers de la llista han estat verificats[212].

Entre els firmants, hi ha un qui és qui en el món de la tecnologia i de la ciència en general, des d'empresaris del sector, com Elon Musk, Steve Wozniak, Emad Mostaque (CEO d'Stability AI), Evan Sharp (cofundador de Pinterest) o Jeff Dean (Google Senior Fellow, investigador de sistemes informàtics a gran escala i sistemes d'IA); fins als divulgadors científics Yuval Noah Harari (historiador famós pel seu llibre *Sapiens*) o Gary Marcus (investigador i divulgador sobre IA); o els professors John J Hopfield (inventor, el 1982, d'una de les primeres xarxes neuronals) o Ramon Lopez De Mantaras; i altes personalitats com Andrew Yang (empresari que es va postular per representar el Partit Demòcrata a les presidencials de 2020, amb propostes com una renda universal de mil dòlars[213]) o George K Rasley Jr (editor de la revista ConservativeHQ.com i antic assistent especial del vicepresident dels Estats Units i membre del personal del Senat i la Cambra de Representants dels Estats Units). Tampoc hi falten ni Jaan Tallin, per part del Centre for the Study of Existential Risk, ni Max Tegmark, per part del Future of Life Institute. A part de les dues entitats mencionades a la línia anterior, també s'hi troben personal de, per exemple, les següents: The Future Society, Berkeley Existential Risk Initiative, Leverhulme Centre for the Future of Intelligence, Foresight Institute o Next Wave Institute. De fet, aquesta vegada sembla que en lloc de

210 FLI (22.03.2023). "Pause Giant AI Experiments: An Open Letter" en *Futur of Life Institute*. Consultat el 14 de juliol de 2023 a https://futureoflife.org/open-letter/pause-giant-ai-experiments/

211 FLI (31.03.2023). *"*FAQs about FLI's Open Letter Calling for a Pause on Giant AI Experiments*" en Future of Life Institute*, pregunta nº10. Consultat el 13 de juliol de 2023 a: https://futureoflife.org/ai/faqs-about-flis-open-letter-calling-for-a-pause-on-giant-ai-experiments/

212 *Ibídem*, pregunta nº3.

213 Andrew Yang (11.07.2023). A *Wikipedia*. Consultat el 14 de juliol de 2023 a: https://en.wikipedia.org/w/index.php?title=Andrew_Yang&oldid=1164903969

mirar qui ha firmat (en alguns casos, fins i tot dues vegades[214]), cal buscar qui falta, com, per exemple, Sam Altman (cofundador d'OpenAI), que deu haver detectat l'interès explícit en parar-li el seu departament d'R+D+I. I és que el document sembla escrit específicament contra aquesta empresa: no només es fa referència a afirmacions que Altman havia fet 6 dies abans a *ABC News* (òbviament, només prenent algunes de les seves afirmacions i oblidant-ne d'altres), sinó que també se la vincula, a través d'un article afegit a peu de pàgina, amb guspires d'una IAG.

El document té set paràgrafs (523 paraules) i va acompanyat de cinc notes a peu de pàgina especificant bibliografia consultada. Amb posterioritat, el 12 d'abril de 2023, també s'hi va enllaçar un document més desenvolupat, de 14 pàgines, titulat "Policymaking in the Pause"[215], que detalla les propostes legislatives que caldria implementar durant els sis mesos de pausa proposats. La carta comença amb una afirmació contundent: P81:«AI systems with **human-competitive intelligence** can pose profound risks to society and humanity»[216]. Cal destacar que, tot i que tractar-los com a sistema no humanitza, el fet de mencionar que poden ser una competència sí que ho fa, i, si no s'afegís el terme "intel·ligència", la cosa no passaria de tenir un toc luddita, però al fer-ho, d'alguna forma queda clar pel lector que hi ha una entitat que pot competir intel·lectualment amb els humans. Aquesta afirmació és fonamentada amb dues notes a peu de pàgina.

La primera nota enllaça amb dotze referències bibliogràfiques, entre les que destaquen ells llibres de Nick Bostrom, *Superintelligence;* Stuart Russell, *Human Compatible: Artificial Intelligence and the Problem of Control;* i el de Mark Tegmark, *Life 3.0: Being Human in the Age of Artificial Intelligence*; i el també supervendes Brian Christian, *The Alignment Problem: Machine Learning and human values*; dues d'aquestes obres són les que precisament menciona Gates en la seva entrada al blog del dia anterior, com es veurà més endavant. També s'hi enllaça una sèrie d'articles de diferent rellevància (quatre dels vuit en *preprint*) tenint en compte el nombre de publicacions dels autors –per exemple, Timnit Gebru i Emily M. Bender, i el seu publicitat "On the Dangers of Stochastic Parrots: Can Language Models Be Too Big" (2021)[217], article el títol del qual ja implica un cert grau etològic a l'identificar els LLM amb lloros estocàstics, tot i que la seva tesi sigui parcialment contrària a una etologia digital; Richard Ngo, que té quatre articles publicats al

214 Casualment, s'ha trobat que Javier Contreras Alcántara havia firmat dues vegades.

215 FLI (31.03.2023). *"*Policymaking in the Pause*" en Future of Life Institute*. Consultat el 14 de juliol de 2023 a: https://futureoflife.org/wp-content/uploads/2023/04/FLI_Policymaking_In_The_Pause.pdf

216 FLI (22.03.2023). "Pause Giant AI Experiments: An Open Letter" en *Futur of Life Institute*, §1. Consultat el 14 de juliol de 2023 a https://futureoflife.org/open-letter/pause-giant-ai-experiments/

217 Aquest article va costar l'acomiadament (o acceptació de renúncia, depenent de qui ho expliqui) de Gebru de Google. Curiosament, Jeff Dean, signant de la carta, fou el responsable de la sortida de Gebru.

2020, el més citat dels quals amb 34 citacions[218]; o Shiri Dori-Hacohen, l'article més citat de la qual té 70 citacions i la suma total dels seus 24 articles, 367[219]. La segona nota enllaça amb l'entrevista a Sam Altman mencionada en el paràgraf anterior. La següent referència de la carta és als Principis d'Asilomar, document que s'ha analitzat com a segona carta oberta. El paràgraf acaba lamentant que tot i aquestes referències aportades, la cursa entre els laboratoris de IA s'hagi desbocat ("out-of-control race") amb l'objectiu de P82:«develop and deploy ever more powerful **digital minds** that no one – not even their creators – can understand, predict, or reliably control»[220]. Aquí s'observa com la identificació d'una ment digital incontrolable pels seus propis creadors (se sobreentén que aquests són humans i que tenen una ment no digital), trasllada al lector la imatge d'un nou ésser, un ésser amb un ment digital, incomprensible i impredictible, trets, tots ells que caracteritzen una etologia digital.

El segon paràgraf segueix amb la mateixa estratègia: identifica els sistemes digitals com a possibles competidors generals dels humans («**human-competitive** at general tasks»[221]), vincula aquestes afirmacions amb bibliografia existent (novament d'OpenAI i concretament, GPT) i encadena una sèrie de preguntes retòriques de caràcter catastrofista que antagonitzen, de vegades "les màquines" o "les ments no humanes", amb un "nosaltres" o "la nostra civilització"; preguntes que són respostes amb una afirmació novament contundent: P83:«Such decisions must not be delegated to unelected tech leaders»[222], afirmació curiosa en un document signat per bona part d'aquest col·lectiu. El paràgraf acaba amb una altra citació d'unes paraules de Sam Altman publicades a la pàgina d'OpenAI amb les quals coincideixen els autors de la carta i que reclamen que se sigui conseqüent amb elles, cosa que implica, segons ells, el que serà la petició general de la carta i que apareix explícitament en el tercer paràgraf: P84:«we call on all AI labs to immediately pause for at least 6 months the training of AI systems more powerful than GPT-4»[223]. Al quart paràgraf, s'apel·la, a través de la quarta nota a peu de pàgina, als Principis en IA de la OCDE (publicats al maig de 2019) i es torna a apel·lar a la impredictibilitat d'aquest models tancats

218 Richard Ngo. A *Google Scholar*. Consultat el 16 de juliol de 2023 a: https://scholar.google.com/citations?user=7CY93A4AAAAJ&hl=en

219 Shiri Dori-Hacohen. A *Google Scholar*. Consultat el 16 de juliol de 2023 a: https://scholar.google.com/citations?hl=en&user=pW5kwa8AAAAJ&view_op=list_works

220 FLI (22.03.2023). "Pause Giant AI Experiments: An Open Letter" en *Futur of Life Institute*, §1. Consultat el 16 de juliol de 2023 a https://futureoflife.org/open-letter/pause-giant-ai-experiments/

221 *Ibídem,* §2.

222 *Ídem*.

223 *Ibídem,* §3.

("black-box") i a les seves capacitats emergents («**emergent** capabilities»[224]). Com ja s'ha dit del verb "aixecar" (*rise*), el verb "emergir", desvinculat del subjecte que provoca la suposada emergència (els redactors de la carta poden estar jugant amb els dos sentits del terme), trasllada també certa idea d'autonomisme o independència entre el que emergeix (la IA) i aquell qui la veu emergir (els humans), cosa que contribueix novament a un element d'etologia. El cinquè paràgraf encadena una sèrie d'adjectius, alguns dels quals tenen un clar significat tecnològic i no etològic (*robust*, *segurs*, *transparent*); uns altres que poden ser interpretats de forma ambivalent, ja sigui tecnològicament o etològicament (*alineat, acurat, interpretable, confiable*[*trustworthy*]); i un que té un significat clarament etològic (*lleial*), doncs lleials ho són les persones (o els gossos), en qualsevol cas, un ésser viu, no una eina (no té cap sentit preguntar-se si un martell és lleial, o un full de càlcul, una pàgina web o un sistema expert, i de fet no és un terme que tingui un ús específic a nivell tecnològic). També hi ha nous usos amb cert caràcter etològic en el sisè paràgraf: P85:«a robust auditing and certification **ecosystem**»[225], concepte associat a la biologia; i un ús de la IA com a subjecte causal explícit en P86:«and well-resourced institutions for coping with the dramatic economic and political disruptions (especially to democracy) that **AI will cause**»[226], frase que deixa clar, per si havia quedat algun dubte, el motiu de fons del text: aconseguir més inversió. El setè i últim paràgraf, tot i l'aparent canvi de to (P87:«Humanity can enjoy a flourishing future with AI»[227]), escateix quina és la visió general de la carta al demanar temps, P88:«and give society a chance to **adapt**»[228], és a dir, ara ja no es tracta de fer eines que s'adaptin a les persones, sinó que són les persones qui s'han d'adaptar a les eines (i per això necessiten un temps), expressió que constitueix el màxim exponent i el grau més perillós d'una etologia digital.

A diferència de la resta de cartes, aquesta ha requerit d'una pàgina de preguntes freqüents específica (FAQs) a causa del nivell de publicitat que va tenir, tal i com admeten en aquesta mateixa entrada (Pregunta 2). També s'hi annexa un document amb recomanacions de caràcter polític-normatiu *(policy recommendations*) de 14 pàgines i en el qual es diferencien set propostes concretes, com la necessitat d'auditar i certificar els sistemes amb IA, regular l'accés d'organització a la potència de càlcul, establir agències estatals d'IA, establir responsabilitat per danys causats per l'IA, introduir mesures per prevenir i rastrejar les filtracions de models d'IA, ampliar el finançament de la investigació tècnica en seguretat de l'IA i desenvolupar estàndards per identificar i gestionar

224 *Ibídem,* §4.

225 *Ibídem,* §6

226 *Ídem.*

227 *Ibídem,* §7.

228 *Ídem.*

contingut i recomanacions generades per l'IA. Tant un document com un altre, tot i incloure algunes expressions etològiques com P89:«AI systems are **growing** ever more powerful»[229], o que els sistemes amb IA P90:«can **learn** and **adapt** after they are sold»[230] i a part de mencionar el suïcidi d'un home després de parlar amb un clone de GPT[231], és molt més curós i específic en la majoria de mencions i evita usos etològics.

Així doncs, de les quatre cartes, està clar que aquesta és la que contribueix més explícitament a una etologia digital, per començar, perquè, a diferència de la primera, els elements etològics apareixen a la carta oberta, document de fàcil accés i publicitat, i no en el document tècnic associat. També ho és més que la segona i la tercera, aquí per la diferència de publicitat i ressò que ha rebut aquest document. Per tot això, si s'intentés representar el nivell de col·laboració en una etologia digital d'aquesta quarta carta, caldria valorar doncs que és un discurs altament potenciat pels mitjans de comunicació, especialment pels afins a Musk, i al qual han donat suport, per algun motiu o altre, moltes personalitats, alguns dels quals amb molta rellevància i influència a nivell tècnic.

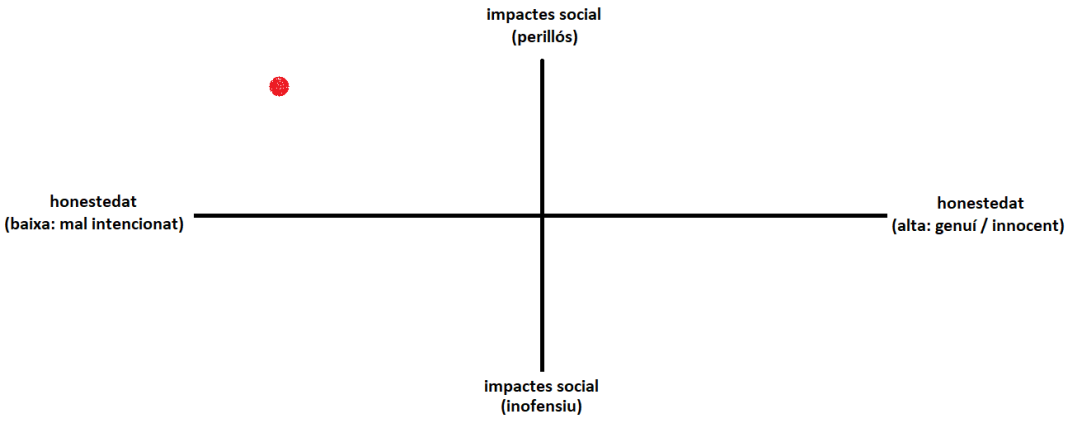

Figura 6: Nivell d'honestedat / impacte social de Musk et al.

229 FLI (31.03.2023). *"*FAQs about FLI's Open Letter Calling for a Pause on Giant AI Experiments*"* en *Future of Life Institute*, pregunta nº5. Consultat el 13 de juliol de 2023 a: https://futureoflife.org/ai/faqs-about-flis-open-letter-calling-for-a-pause-on-giant-ai-experiments/

230 FLI (31.03.2023). *"*Policymaking in the Pause*"* en *Future of Life Institute,* pàg. 9. Consultat el 16 de juliol de 2023 a: https://futureoflife.org/wp-content/uploads/2023/04/FLI_Policymaking_In_The_Pause.pdf

231 Es refereixen a la notícia que va saltar a la premsa el dia 1 d'abril en un diari belga i que relacionava el suïcidi d'una persona amb les converses que havia tingut amb un model GPT-J no vinculat a l'empresa OpenAI. Consultat el 16 de juliol a: https://es.euronews.com/next/2023/04/01/un-hombre-se-suicida-despues-de-que-un-chat-de-ia-le-invitara-a-hacerlo

La taula intenta il·lustrar com un nivell baix d'honestedat o integritat, però recolzada i ampliada per una potent estructura comunicativa, pot tenir un alt impacte social.

3.3 La proposta de Bill Gates

Tres verbs, *inspire-convince-excite*, reflecteixen l'estat d'ànim que transmet el text de Gates. La primera part, sense títol específic a diferència de les altres sis, ubica històricament el lector dins de les coordenades vivencials de Gates: P91:«In my lifetime, I've seen two demonstrations of technology that struck me as revolutionary»[232]. La primera d'aquestes tecnologies és la pantalla gràfica; la segona és GPT, és a dir, la utilització de models de llenguatge gran (LLM) per desenvolupar xatbots. També en aquesta primera part, Gates defineix clarament l'objectiu d'aquesta tecnologia amb una predicció: P92:«It will change the way people work, learn, travel, get health care, and communicate with each other. Entire industries will reorient around it. Businesses will distinguish themselves by how well they use it»[233]. Fins i tot planteja el seu somni, identificant-lo clarament amb la seva tasca filantròpica (a diferència de la tasca empresarial, de la qual sembla desvincular-se o subsumir-la a les estones lliures, en la mesura que la filantròpica és a «full-time job these days»[234]): P93:«I've been thinking a lot about how—in addition to helping people be more productive—AI can reduce some of the world's worst inequities»[235]. Per Gates, la principal inequitat global és sanitària i, en concret, els 5 milions d'infants de menys de cinc anys que moren anualment. Els altres dos pilars de la inequitat són l'educació i el canvi climàtic. Sanitat i educació mereixeran un apartat independent del text, mentre que les problemàtiques del canvi climàtic només tornaran aparèixer com la causa de problemes en països pobres («low-income countries»). En els dos darrers paràgrafs d'aquesta primera part, Gates mostra novament el seu entusiasme davant d'aquesta gran oportunitat que presenta la IA, però també assumeix que aquesta nova tecnologia implicarà una sèrie de canvis disruptius i problemes que caldrà afrontar com ara P94:«the workforce, the legal system, privacy, bias, and more»[236], problemes sobre els quals farà propostes en un apartat específic del text titulat "Risks and problems with AI".

Alguns aspectes que cal destacar d'aquesta primera part és com en tot moment Gates es refereix a la intel·ligència artificial com una tecnologia, equiparable a la pantalla gràfica, els microprocessadors, els ordinadors personals, els telèfons mòbils o internet. Aquesta comparativa,

232 GATES, Bill (21.03.2023). "The Age of AI has begun" en *GatesNotes. The blog of Bill Gates*, §1. Consultat el 6 de juliol de 2023 a: https://www.gatesnotes.com/The-Age-of-AI-Has-Begun

233 *Ibídem,* §9.

234 *Ibídem,* §10.

235 *Ídem*.

236 *Ibídem,* §14.

lluny d'humanitzar l'eina, el que fa és situar-la clarament com un objecte: no hi ha confusió possible entre l'eina i qui la maneja, és a dir, les persones. Les persones són a qui ha de millorar la vida aquesta eina i, tot i que apunta que canviarà la forma de treballar, aprendre o viatjar, no són les persones les que modificaran els seu comportament per adaptar-s'hi (a l'igual que les persones no van adaptar-se a la roda, tot i que l'invent de la roda va canviar la forma de viatjar de les persones). Si es vol, es pot apuntar que hi ha certa ingenuïtat en la pretensió de poder aportar maneres de mitigar els riscs de la IA merament des de la filantropia i en un document de sis pàgines com aquest, però en cap moment es confon una eina tecnològica amb un ésser viu ni s'apliquen verbs associats a comportaments biològics a aquesta eina. Tampoc s'hi fa referència amb un pronom personal (s'utilitza sempre el pronom no-personal «it») i, tot i que s'hi vinculen expressions metafòriques com P95:«GPT got a 5», «it has aced the test», «it wrote a thoughtful answer», «AI can help», cap d'elles en un context que generi confusió possible. L'única ocasió en què s'hi aplica un verb mental és en la següent frase: P96:«AIs also make factual mistakes and experience hallucinations»[237]; tanmateix, el fet d'emfatitzar la propietat factual d'aquest errors (vinculats a un fet extern, objectiu), de vincular-lo al concepte tècnic "al·lucinació" i de fer-ho just abans del segona apartat, titulat "Defining artificial intelligence", no genera en el lector cap confusió d'arrel antropomòrfica.

Aquest segon apartat és una mostra més de com Gates es vol separar de discursos catastrofistes, alguns dels quals vinculats a una etologia digital. Comença fent una diferència entre IA i IAG i deixant clar què pot fer ChatGPT i què no pot fer: P97:«It is learning how to do chat better but can't learn other tasks»[238]. També emfatitza que encara no existeix una IAG i que és un tema en debat dins la indústria de la computació si és possible crear-les. De fet, implícitament Gates sembla reconèixer que, arran de l'aprenentatge automàtic i de la potència actual de càlcul, ja hi ha aspectes, a part del càlcul, en els quals aquestes eines són millor que els humans i que, ràpidament, encara milloraran més. Tanmateix, aquesta insinuació, en el context del següent apartat –en el qual afirma explícitament que els humans encara són millor que GPT en moltes coses malgrat que aquestes no siguin útils en diverses feines–, no sembla tan encaminada en minoritzar els humans, sinó en justificar per què hi ha feines en les quals una IA pot millorar el rendiment humà. Aquest és l'objectiu del tercer apartat, titulat "Productivity enhancement": P98:«[...] these data sets [referint-se a les dades que utilitzen les empreses per formar als seus treballadors] will also be used to train the AIs that will empower people to do this work more efficiently»[239]. Malgrat haver afirmat anteriorment que el seu interès a temps complet és filantròpic, a ningú se li escapa que Gates és

237 *Ídem.*

238 *Ibídem,* §15.

239 *Ibídem,* §18.

cofundador de Microsoft, empresa comercialitzadora de software i serveis de productivitat i processos de negoci. De fet, és únicament en aquest context en el qual s'antropomorfitza les possibilitats de GPT, el qual se li atorguen diferents rols: copilot, agent personal, assistent personal digital i, comparativament, P99:«like having a white-collar worker available to help you with various tasks»[240]. També se li atribueixen diferents verbs relacionats amb activitats mentals com: habilitat d'expressar idees (§19), comprovar [els últims correus] (§21), conèixer [les reunions que hom atén] (§21), llegir [el que hom llegeix] (§21), llegir [el que hom no li plau llegir] (§21). Precisament, aquest és l'argument principal: P100:«This will both improve your work on the tasks you want to do and free you from the ones you don't want to do»[241]. Per tant, queda clar que les tasques encomanades a GPT són d'un perfil més baix de les que pot fer una persona (en una feina d'oficina, se sobreentén), i que aquestes estan relacionades en la presa de decisions repetitives sense un valor aparent. Per altra banda, també es reconeix que utilitzar aquesta tecnologia així no es possible encara, però Gates defensa que gràcies els avanços fets en IA, ara és un objectiu realista.

L'únic moment en tot el text en el qual Gates s'aproxima a una etologia digital és també en aquest context empresarial, quan imagina com seria treballar amb una eina així:

> P101: Company-wide agents will empower employees in new ways. An agent that understands a particular company will be available for its employees to consult directly and should be part of every meeting so it can answer questions. It can be told to be passive or encouraged to speak up if it has some insight. It will need access to the sales, support, finance, product schedules, and text related to the company. It should read news related to the industry the company is in. I believe that the result will be that employees will become more productive.[242]

En aquest paràgraf Gates clarament antropomorfitza el ChatGPT: se l'imagina pràcticament de cos present en les reunions d'empresa, com el cap de Jeremy Bentham en les reunions del Consell de la University College de Londres. Tanmateix, el paràgraf queda molt circumscrit en un àmbit laboral, com a suport als treballadors i en la millora general de la humanitat, com expressa clarament en el següent paràgraf, en el qual, tot i utilitzar una expressió pròpia dels autors catastrofistes ("the rise of", associada a l'alliberament d'algú oprimit), li dona el sentit contrari: P102:«The rise of AI will free people up to do things that software never will—teaching, caring for patients, and supporting the elderly, for example»[243]. Per tant, queda clar per on van les expectatives

240 *Ibídem,* §19.

241 *Ibídem,* §21.

242 *Ibídem,* §23.

243 *Ibídem,* §24.

de Gates i que aquestes tenen un caràcter merament utilitari, és a dir, la IA és una eina al servei dels humans i, en cap cas, una nova espècie.

Això queda també de manifest en els següents dos apartats, "Health" i "Education", on Gates intenta buscar algun sentit a la utilització, primer del GPT, i després, a la resta d'eines d'IA. Tal i com ha reconegut en el paràgraf 24, hi ha camps que difícilment podran ser ocupats per eines d'aquest tipus, però això no desanima un Gates que, tot i els seus interessos filantròpics (cal recordar que l'objectiu final és reduir la inequitat mundial), no deixa de voler trobar alguna utilitat a les eines que comercialitza. Pel que fa al GPT, Gates creu que pot ser útil per a tasques burocràtiques relacionades tant amb la medicina com l'educació, com poden ser la complementació de reclamacions d'assegurances i l'elaboració d'esborranys de les notes de visita d'un metge, o l'avaluació de la comprensió de l'alumnat i l'aportació de consells en plans de carrera respectivament. En el cas de medicina, destaca l'ús que es podria fer en el triatge bàsic de pacients i també donant consell sobre si uns símptomes requereixen o no d'un tractament mèdic, és a dir, aspectes que exigeixen la supervisió d'un professional de carn i ossos. En el cas d'educació, tot i reconèixer que la implementació de la computació en aquest camp no ha tingut els efectes desitjats –de fet, reconeix que un dels pocs èxits en aquest camp ha estat la Viquipèdia (§35)–, no s'està de predir que en cinc o deu anys aquest software acomplirà la promesa de revolucionar la manera com s'ensenya i s'aprèn: P103:«It will know your interests and your learning style so it can tailor content that will keep you engaged. It will measure your understanding, notice when you're losing interest, and understand what kind of motivation you respond to. It will give immediate feedback»[244]. Com que aquí no s'analitzen ni les conseqüències d'aquesta educació monitoritzada –talment cada alumne fos un pacient del qual es llegeixen i s'apunten les reaccions vitals per tal de controlar-les i, si cal, modificar-les– ni tampoc es valora la bondat de la proposta, només s'observarà que un sistema com el que proposa, si es vol integrar de forma definitiva, va més en la línia del transhumanisme que no pas de l'etologia digital.

En qualsevol cas, ni en l'apartat de salut ni en el d'educació hi ha cap proposició que contingui elements d'una etologia digital i els respectius paràgrafs semblen una promesa d'algú enlluernat per aquesta tecnologia i de la qual té molt interès en treure'n rendiment. De fet, tant en el cas de la medicina –quan afirma que P104:«Some companies are working on cancer drugs that were developed this way»[245]–, o en el cas de l'educació –quan afirma P105:«AIs will need a lot of training and further development before they can do things like understand how a certain student

244 *Ibídem,* §36.

245 *Ibídem,* §32.

learns best or what motivates them»[246] o directament quan diu que P106:«New tools will be created for schools that can afford to buy them, but we need to ensure that they are also created for and available to low-income schools in the U.S. and around the world»[247]–, sembla que Gates està predient 1) com es pot monetitzar la inversió feta («cancer drugs» i «schools than can afford»), 2) la inversió pendent («need a lot of training» no deixa de ser un eufemisme per reclamar més inversió), i 3) per quan serà possible un retorn de la inversió («to low-income schools in the U.S. and around the world» és equivalent a dir quan s'implementi també a les escoles públiques a través de la compra per part dels respectius governs).

Després d'abordar què pot fer la IA per salut i educació, Gates afronta els riscs i problemes amb la IA en el sisè apartat d'aquest text. Per una banda, hi ha el que ell considera problemes tècnics resolubles amb temps («in less than two years and possibly much faster»[248]), com poden ser la dificultat d'un LLM per entendre el context o errors matemàtics deguts a dificultats amb el raonament abstracte. Per una altra, hi ha les preocupacions no tècniques, com l'amenaça que poden suposar si aquestes eines són utilitzades amb fins deshonestos, cosa que requeriria que els governs treballessin amb el sector privat per limitar aquests riscos (§43). I, finalment, Gates entra a comentar les possibilitats que es creï una IA forta o general que escapi del control humà i decideixi que els humans són una amenaça. La postura de Gates davant d'aquesta hipòtesi és explícitament agnòstica i implícitament escèptica; agnòstica quan afirma: P107:«But none of the breakthroughs of the past few months have moved us substantially closer to strong AI. Artificial intelligence still doesn't control the physical world and can't establish its own goals»[249]; escèptica quan afirma: P108:«Once developers can generalize a learning algorithm and run it at the speed of a computer—an accomplishment that could be a decade away or a century away—we'll have an incredibly powerful AGI»[250]. És a dir, sense desinflar la bombolla de les expectatives que pot generar la IA (de fet, afirma també P109:«Superintelligent AIs are in our future»[251]), expectatives que generen negoci, Gates deixa clar que GPT no té res a veure amb aquests relats apocalíptics i que els LLM no han implicat un canvi en les perspectives d'aparició d'una IAG. De fet, el seu elogi als autors que viuen de generar aquests discursos de la por (Gates menciona tres exemples *Superintelligence*, de Nick

246 *Ibídem,* §38.

247 *Ibídem,* §39.

248 *Ibídem,* §42.

249 *Ibídem,* §47.

250 *Ibídem,* §45.

251 *Ídem.*

Bostrom; *Life 3.0*, de Max Tegmark; i *A Thousand Brains*, de Jeff Hawkins) sembla sobretot un escarni: «all three books are well written and thought-provoking»[252].

L'últim apartat del text es titula "The next frontiers" i Gates acaba de desplegar la seva visió: P110:«There will be an explosion of companies working on new uses of AI as well as ways to improve the technology itself»[253], P111:«There will be immense competition on both approaches»[254] (referint-se a si és millor desenvolupar algoritmes especialitzats o alguns de caràcter més general), P112:«No matter what, the subject of AIs will dominate the public discussion for the foreseeable future»[255]. Sembla acurat dir que Gates pretén seguir generant expectatives: una explosió de companyies que competiran entre elles per l'imminent negoci de la integració de la IA en tots els sectors. També sembla acurat dir que la seva predicció és optimista i que traspua il·lusió i genuïnitat, i també un pel de ingenuïtat, especialment quan vincula el desenvolupament d'aquesta tecnologia amb la millora de la vida de les persones. De fet, els tres principis que proposa per abordar aquest nou panorama semblen provenir del manual del bon viatjant: equilibrar pors i esperances, repartir-ne els beneficis (cosa que implica també socialitzar-ne les pèrdues) als sectors més vulnerables per reduir la inequitat mundial, i recordar que les limitacions actuals només són temporals (§55). Resulta interessant, perquè encaixa amb la imatge que transmet Gates en aquest text, l'última reflexió/elucubració/desig/somni: P113:«it's interesting to think about whether artificial intelligence would ever identify inequity and try to reduce it. Do you need to have a sense of morality in order to see inequity, or would a purely rational AI also see it?»[256]. Aquí és podria plantejar si no hi ha certs detalls propis d'una etologia digital, ara bé, Gates separa clarament allò que constitueix una persona (ésser racional amb un sentit de la moralitat) del que realment es pot digitalitzar (les estructures racionals), cosa que l'allunya, novament, d'aquesta etologia digital.

Per acabar, Gates reclama novament certes normes clares («the world needs to establish the rules of the road»[257]) i algun tipus de regulació que redueixi els riscos inherents a qualsevol nou invent, tal i com també faran Musk i companyia en la seva carta, publicada l'endemà d'aquesta entrada amb un estil completament diferent al d'aquest text com s'ha vist.

Recapitulant, el text de Gates, tot i que no és innocent, doncs prové d'una persona amb uns forts interessos econòmics en el tema (cal recordar que el gener d'aquest any Microsoft va fer una

252 *Ibídem,* §48.

253 *Ibídem,* §49.

254 *Ibídem,* §50.

255 *Ibídem,* §51.

256 *Ibídem,* §55.

257 *Ibídem,* §56.

nova inversió en OpenAI, ara de 10.000 milions de dòlars[258] després dels 1.000 milions aportats entre 2019 i 2021[259]), és honest i va de cara, per això podria titllar-se d'inofensiu des d'un punt de vista de l'etologia digital: no hi ha una pretensió real de fer passar un ens digital per un ésser viu. En aquest sentit, es pot considerar un text sense mala intenció ni perill. Ara bé, per l'altra banda, està clar que el projecte de Gates és, tard o d'hora, acabar aconseguint algun tipus d'IA que realment sigui una IA, és a dir, una entitat pròpia, i no és casualitat que se la imagini present i personificada a les reunions, com un consultor expert fidel al seu costat.

Per tant, la seva representació gràfica podria il·lustrar-se així:

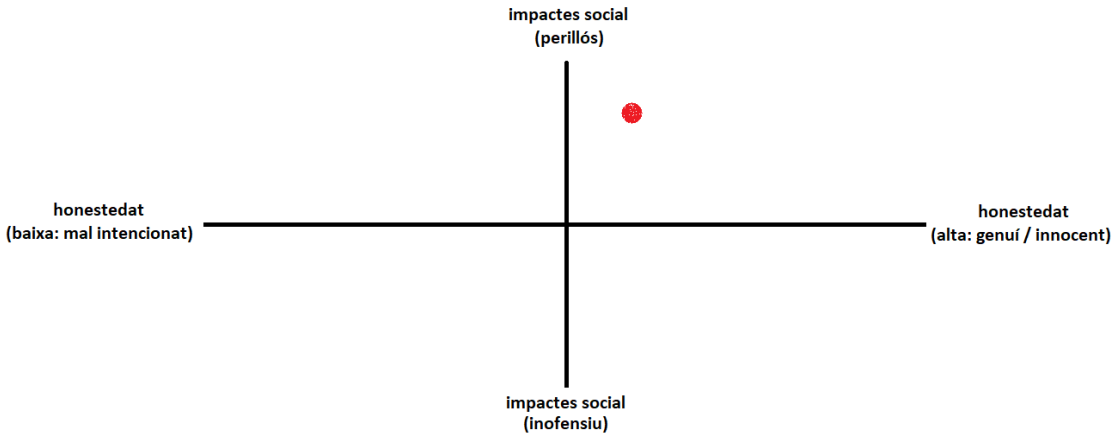

Figura 7: Nivell d'honestedat / impacte social de Bill Gates

La imatge intenta representar com, tot i puntuar positivament en honestedat, tampoc se li atorga la màxima puntuació, ja que té interessos econòmics que sempre poden influir, encara que sigui inconscientment, en la forma de pensar. I al mateix temps, que la certa coherència del seu discurs i, sobretot, la gran capacitat operativa que té al darrere, fan que aquest pugui tenir un impacte alt a nivell social.

3.4 Conclusions

L'estudi d'aquests documents, per una banda la proposta de Musk *et alii* i per l'altra la de Gates, ha servit per presentar les diferents maneres en què es pot portar a terme un mateix objectiu, això és, reclamar més inversió. Aquí no s'ha volgut tractar explícitament si els discursos promoguts

258 *Forbes* (27.01.2023). "Microsoft Confirms Its $10 Billion Investment Into ChatGPT, Changing How Microsoft Competes With Google, Apple And Other Tech Giants" en *Forbes*. Consultat el 12 de juliol de 2023 a https://www.forbes.com/sites/qai/2023/01/27/microsoft-confirms-its-10-billion-investment-into-chatgpt-changing-how-microsoft-competes-with-google-apple-and-other-tech-giants/

259 *OpenAi* (22.07.2019). "Microsoft invests in and partners with OpenAI to support us building beneficial AGI" en *OpenAi.* Consultat el 12 de juliol a https://openai.com/blog/microsoft-invests-in-and-partners-with-openai

per aquestes persones i també entitats és més o menys encertat, sinó si les seves paraules promouen o no una etologia digital i en quin grau ho fan. Com s'ha vist, la proposta de Gates no contribueix pràcticament gens a promocionar una etologia digital, mentre que la de Musk *et alii* sí, tot i que de forma desigual. Entre la primera carta (2015) i l'última (2023) hi ha hagut un canvi de sentit en la forma de tractar els sistemes amb IA, des d'una visió més optimista i menys etològica (si només es té en compte el text de la primera carta i no els últims paràgrafs del document annexat), fins a una visió més catastrofista i més etològica (si només es té en compte el text de la quarta carta i no els documents annexats). Aquest canvi pot respondre al grau d'influència d'autors més propensos a discursos alarmistes o exagerats (Bostrom, Chalmers, Russell, Tegmark o Christian) sobre el conjunt d'interessats, així com al fet que els tres primers documents s'escriuen abans de que es doni accés a ChatGPT, primera mostra pública de què poden fer aquests tipus de LLM. També ha canviat al grau d'influència dels seus protagonistes en els mitjans de comunicació: és obvi que la rellevància de Musk a nivell de mercat de l'opinió pública ha augmentat molt, especialment després de la compra de Twitter i algunes de les seves declaracions més incendiàries.

En qualsevol cas, l'anàlisi d'aquestes cartes ha servit com exemple per mostrar què és una etologia digital i com es pot detectar. També per evidenciar com darrere d'una etologia digital sol haver-hi una petició d'inversió, i que hi ha altres maneres de fer el mateix sense necessitat d'inventar una nova espècie. Tanmateix, sembla que hi ha entitats que viuen precisament d'alimentar aquestes pors, els fundadors de les quals a vegades també són propietaris d'empreses d'avaluació de riscos, entre ells, dels mateixos sobre els que aquí alarmen[260].

260 Per exemple, Anthony Aguirre, firmant de tres de les quatre cartes, treballa tant a Future for Life Institut com és membre de la junta de Metaculus, empresa que es defineix així: «Metaculus offers trustworthy forecasting and modeling infrastructure for forecasters, decision makers, and the public». Consultat el 17 de juliol de 2023 a: https://www.metaculus.com/

Unfortunately, nature seems unaware of our intellectual
need for convenience and unity, and very often takes
delight in complication and diversity

Santiago Ramon y Cajal, *Nobel Lecture*

4. Els autors escèptics

Aquest capítol presenta tres propostes més d'etologia digital, però allunyades de l'estratègia de la por: aquests autors aposten per una investigació honesta i rigorosa, i tenen un objectiu clar que s'assumeix que encara pot estar molt lluny. Aquesta estratègia té un doble vessant: pot no ser tan sorollosa, però és més fecunda a llarg termini, com es veurà amb alguns dels invents que ha patentat l'últim dels autors, Rodney Brooks.

4.1 En contra dels autors de la por

Dins del món de la investigació en intel·ligència artificial, hi ha una sèrie d'autors que rebutgen i critiquen els discursos basats en la por per la seva falta d'objectivitat i per inflar artificialment la bombolla de la IA, valgui la redundància. Entre aquests autors i obres, hi ha Gary Marcus i Ernest Davis i el seu *Rebooting AI. Building Artificial Intelligence We Can Trust*, Melanie Mitchell i *Artificial Intelligence. A Guide for Thinking Humans*, Rodney Allen Brooks i "Elephants Don't Play Chess", entre d'altres. Tots són coneguts per la seva contribució a la investigació en IA i treballen o han treballat llargament per empreses del sector o departaments d'investigació de reputades facultats de tecnologia. És a dir, no són sospitosos de luddisme. Tanmateix, són dels pocs que dins del sector han alçat la veu per alertar del perill del *hype* que es pot estar creant i com pot ser-ne de contraproduent: prometre el que no es pot aconseguir, a llarg termini sol portar a un nou hivern.

Tot i això, resulta complicat saber, únicament a través de l'observació del seu comportament públic i declaracions, quina és exactament la seva posició respecte a la IA en general i a la seva investigació en particular. Per exemple, després d'escriure *Rebooting AI* (2019) –una obra que va ser entesa com «a welcome antidote to the hype that has engulfed AI over the past decade»[261]– i d'afirmar, en una entrada conjunta amb Ernest Davis al seu blog, que el ChatGPT cometia els mateixos errors que els models anteriors[262], Gary Marcus va firmar la "Carta oberta" de Musk i

261 BROOKS, Rodney. Citat en MARCUS, Gary; DAVIS, Ernest (2019). *Rebooting AI. Building Artificial Intelligence We Can Trust*, Vintage books de Penguin Random House, Nova York, 2020.

262 MARCUS, Gary; DAVIS, Ernest (10.01.2023). "Large Language Models like ChatGPT say The Darnedest Things" en *BLOG@CACM*. Consultat el 11 d'agost de 2023 a: https://cacm.acm.org/blogs/blog-cacm/268575-large-

companyia i va declarar que el dia que va ser cridat a declarar davant del senat sobre la necessitat de regular la investigació en IA, va ser «the highlight of my career»[263]. També Yann LeCun, el 26 de setembre de 2019, havia firmat un article conjuntament amb Anthony Zador, titulat "Don't Fear the Terminator", en el qual defensava que una mala comprensió de la IA era la causa de distracció de «more mundane but far more likely posed by the technology in the near future».[264]

La fluctuació aparent de les opinions d'aquests investigadors, que tot i coneixent l'estat de la qüestió poden defensar un discurs el dia que treballen a una universitat i un de diferent el dia que els contracten a una multinacional (o simplement, quan reben finançament d'una entitat o d'una altra), no permet concloure si aquestes personalitats participen sempre conscientment en la confecció d'una etologia digital, però sí que permet veure fins a quin punt en cert text hi poden col·laborar més o menys. Per tant, aquest apartat se centrarà en intentar analitzar tres posicions escèptiques, però amb matisos diferents. Per una banda, la posició de Gary Marcus i Ernest Davis en el seu *Rebooting AI* i també el butlletí incombustible de Marcus. Després, els documents acadèmics d'una investigadora com Melanie Mitchell qui no ha patit tantes vacil·lacions: no ha firmat la carta i, de fet, ha escrit explícitament en contra d'ella. Finalment, la posició de Rodney Brooks en "Elephants Don't Play Chess" i com aquesta va canviant fins a la que defensa en algunes de les entrades del seu blog o a "The Cul-de-Sac of the Computational Metaphor" (aquest canvi servirà per veure el viratge entre el primer Brooks i el segon Brooks, qui s'utilitza com exemple de que una altra manera d'encarar la investigació és possible).

4.2 Reiniciant Marcus

Gary Marcus (1970) es descriu a ell mateix com un de les veus cantants (*leading voice*) en IA, científic, autor supervendes i emprenedor en sèrie (fundador de Robust.AI i Geometric, empresa adquirida el 2016 per Uber)[265]. És doctor en Ciències cognitives pel Massachusetts Institute of Technology, professor emèrit de Psicologia a la New York University i fou director del Center for

language-models-like-chatgpt-say-the-darnedest-things

263 MARCUS, Gary (08.06.2023). "Two models of AI oversight – and how things could go deeply wrong" en *Marcus on AI*. Consultat el 02 d'agost de 2023 a: https://garymarcus.substack.com/p/two-models-of-ai-oversight-and-how. Després d'aquestes afirmacions en la seva *newsletter*, Marcus, tot i tenir una mica de remordiments, ha seguit justificant la seva firma. En el butlletí de 21 de setembre de 2023, es defensa d'haver promogut el *hype* i assumeix que la carta de Musk i companyia només anava contra OpenAI: https://garymarcus.substack.com/p/six-months-after-the-pause-letter

264 ZADOR, Anthony; LECUN, Yann (26.092019). "Don't Fear the Terminator" en *Scientific American*. Consultat el 19 de juliol de 2023 a: https://blogs.scientificamerican.com/observations/dont-fear-the-terminator/

265 MARCUS, Gary (2022). *Gary Marcus*. Consultat el 14 de juliol de 2024 a: http://garymarcus.com/

Child Language de la mateixa NYU[266]. La seva visió de la IA és des de l'estudi psicològic de la intel·ligència i defensa una visió nativista de la ment, concretament, com un nyap (Marcus utilitza el l'argot *kluge* per referir-s'hi, que en informàtica descriu un programa construït a partir de correccions i trossos d'altre codi, també conegut com a pedaç o *parche*). En canvi, Ernest Davis, professor de Ciències de la computació també a la NYU[267], és especialista en el tractament computacional del llenguatge per recrear el sentit comú, com es pot veure en la relació d'articles publicats[268]. Mentre que Davis és un home discret que només apareix als mitjans com a coautor amb Marcus, aquest és conegut per les seves habituals aportacions a les polèmiques més acarnissades de Twitter i, des del 14 de maig de 2022, un contribuïdor constant a Substack, eina per crear butlletins de notícies (*newsletters*): ha enviat 228 butlletins entre llavors i el 14 de juliol de 2024, és a dir, 8,7 butlletins al mes[269]. Tenint en compte aquesta prolífica dedicació, sorprèn com encara ha tingut temps per escriure *Taming Silicon Valley. How we can ensure that AI works for us*, previst de publicar-se el 24 de setembre de 2024[270], i que, segons ha promès en el butlletí del 27 d'abril de 2024[271], tractarà del comportament dels dirigents de les grans companyies d'informàtica (cosa que pot fer la competència a aquest treball).

Sense excuses: tècnics i coneixedors del producte

Marcus i Davis coneixen bé el funcionament intern dels seus respectius productes: Marcus coneix i entén tot el que tècnicament se sap avui en dia del funcionament de la ment humana, cosa que li permet afirmar coses com la següent:

> P114 «The behavior of machines is often superficially similar to the behavior of humans, so we are quick to attribute to machines the same sort of underlying mechanisms, even when they lack them».[272]

266 "Research affiliates" en *Center For The Study Of Human Origins*. Consultat el 14 de juliol de 2024 a: https://wp.nyu.edu/csho/people/affiliated_researchers/

267 DAVIS, Ernest. *New York University*. Consultat el 14 de juliol de 2024 a: https://cs.nyu.edu/~davise/index.html

268 DAVIS, Ernest (2021). "Research Papers" en *New York University*. Consultat el 14 de juliol de 2024 a: https://cs.nyu.edu/~davise/pubs.html

269 "Gary Marcus" en *Substack*. Consultat el 14 de juliol de 2024 a: https://substack.com/@garymarcus

270 "Taming Silicon Valley" en *Penguin Random House*. Consultat el 14 de juliol de 2024 a: https://www.penguinrandomhouse.com/books/768076/taming-silicon-valley-by-gary-f-marcus/

271 MARCUS, Gary (27.04.2024). "We can't trust the fox to guard the henhouse, especially when it comes to AI" en *Marcus on AI*. Consultat el 14 de juliol de 2024 a: https://garymarcus.substack.com/p/we-cant-trust-the-fox-to-guard-the

272 MARCUS, Gary; DAVIS, Ernest (2019). *Rebooting AI. Building Artificial Intelligence We Can Trust*, Nova York, Vintage Books (Penguion Random House LLC), 2020, pàg. 18.

Per part de Davis, té un coneixement tècnic detallat del funcionament intern dels sistemes d'aprenentatge automàtic i aprenentatge profund, però també de com s'havia tractat el problema en l'època de la GOFAI (*Good Old-Fashion Artificial Intelligence*), és a dir, de quan es prioritzava l'aproximació simbòlica com la del seu mentor Drew McDermott, autor d'un dels primers documents crítics amb el *hype*, "Artificial intelligence meets natural stupidity" (1976).

> P115 «Crucially, AI is not magic, but rather just a set of engineering techniques and algorithms, each with its own strengths and weaknesses, suitable for some problems but not others».[273]

Per tant, en el seu text hi ha un esforç explícit per no enganyar al públic amb pors infundades, mecanisme que s'havia vist associat a la confecció d'una etologia digital intuïtiva:

> P116 «AI doesn't have to want to destroy us in order to create havoc. In the short term, what we should worry most about is whether machines are actually capable of reliably doing the tasks that we assign them to do».[274]

El seu objectiu és el d'explicar amb pèls i senyals el problema actual de la IA generativa:

> P117 «As we will discuss later, deep learning systems in no way capture the complexity and diversity of actual brains and the components of deep learning systems lack virtually all the complexity of actual neurons. As the late Francis Crick noted, it's a serious stretch to call them brain-like».[275]

Tanmateix, és en assenyalar aquest problema que Marcus i Davis comencen a construir, incipientment, una etologia digital, com es pot apreciar en P117 quan s'identifica l'objecte al qual cal assimilar-se (o simular), el cervell humà; i això, inevitablement, porta a fer comparacions:

> P118 «For real intelligence you also need reasoning, language, and analogy, none of which is nearly so well handled by current technology».[276]

> P119 «We need systems that can truly reason about the complex interplay of entities that causally relate to one another in an ever-changing world».[277]

> P120 «The machine-reading systems of our dreams, when they arrive, would be able to answer essentially any reasonable question about what they've read […]. Just as a college

273 *Ibídem*, pàg. 24.

274 *Ibídem*, pàg. 30.

275 *Ibídem*, pàg. 42. Nota al peu per aclarir el concepte de "neural network".

276 *Ibídem*, pàg. 64.

277 *Ibídem*, pàg. 66.

student writing a term paper can bring together ideas from multiple sources, cross-validating them and reaching novel conclusions, so too should any machine that can read».[278]

P121 «Then finally the keystone: construct a kind of human-inspired learning system that uses all the knowledge and cognitive abilities that the AI has; that incorporates what it learns into its prior knowledge; and that, like a child, voraciously learns from every possible source of information: interacting with the world, interacting with people, reading, watching videos, even being explicitly taught. Put all that together, and that's how you get to deep understanding».[279]

De P118 a P121 hi ha una gradació creixent que acaba assumint les primeres forma explícites d'una etologia digital: primer, en P118 i P119 l'apropiació de funcionalitats (raonament, llenguatge i analogia) i que mantinguin una relació causal amb el món; en P120, ja hi ha una equivalència amb un estudiant universitari; i, finalment, en P121, una comparació amb un nen. Tots aquests trets han estat identificats com elements propis d'una etologia digital intuïtiva.

El gir marcusià

Segurament, el següent és el paràgraf que descriu millor aquest canvi de paradigma, que aquí s'identifica amb el gir marcusià que permet passar d'una etologia digital intuïtiva a una etologia digital raonada, fonamentada, i que irònicament en aquest treball s'identifica amb aquest reinici que proposa Marcus:

P122 «Computers don't have to work in the same ways as people. There is no need for them to make the many cognitive errors that impair human thought, such as confirmation bias (ignoring data that runs against your prior theories), or to mirror the many limitations of the human mind, such as the difficulty that human beings have in memorizing a list of more than about seven items. There is no reason for machines to do math in the error-prone ways that people do. Humans are flawed in many ways, and machines need not inherit the same limitations. All the same, there is much to be learned from how human minds—which still far outstrip machines when it comes to reading and flexible thinking—work».[280]

Per una banda, s'assumeix que els ordinadors, com qualsevol eina, no han de per què assimilar-se als humans, sinó que poden fer tasques específiques sense cometre els mateixos errors o millorant-ne les qualitats (com un martell pica més fort que un puny o una excavadora permet fer

278 *Ibídem*, pàg. 70-71.

279 *Ibídem*, pàg. 179.

280 *Ibídem*, pàg. 118.

forats més ràpidament que un parell de mans). Però, al mateix temps, considera que el mirall és l'ésser humà, que cal inspirar-s'hi, cosa que tard o d'hora, però inevitablement, acaba portant a una inversió de la metàfora computacional:

> P123 «(As he noted there, each gene is something like an "IF-THEN" statement in a computer program. The THEN side specifies a particular protein to be built, but that protein is only built IF certain chemical signals are available, with each gene having its own unique IF conditions. The result is like an adaptive yet highly compressed set of computer programs, executed autonomously by individual cells, in response to their environments. Learning itself emerges from this stew)».[281]

Aquí, en un especificació secundària entre parèntesis, ja no és que una estructura condicional programada simuli un procés natural, sinó que per entendre com funciona una seqüència natural, el gen, es requerix pensar-lo com un tros de codi informàtic (ja s'ha observat anteriorment la coincidència en la referència al terme *codi*), és a dir, és el gen el que és comparable amb el codi. És cert que aquesta inversió en l'ordre de la metàfora no és un recurs àmpliament utilitzat per Marcus i és cert també que ell no n'és l'inventor i que ho fa en una part molt secundària i entre parèntesis de l'argumentació, però no deixa de ser interessant (si més no psicològicament) que hi caigui, ja que, tot i la cura que dedica Marcus a evitar aquestes imprecisions, en cert moment se li escapa una, cosa que denota que la metàfora computacional o, en aquest cas, la seva inversió, està plenament assentada en l'imaginari col·lectiu.

Hi ha un altre argument de Marcus en el que també sembla aparèixer, a contracor quasi bé, un detall propi d'una etologia digital i que, conceptualment, coincideix amb un passatge citat anteriorment de Frans de Waal, reputat etòleg. Tal i com s'ha vist en el capítol 2, de Waal fa servir un fetge per explicar el procediment científic i, per context, el de l'etologia en concret; aquí Marcus fa servir un ronyons per justificar per què cert reduccionisme tecnològic no és raonable:

> P124 «In a sort of terminological imperialism, advocates of deep learning often refer to a system, no matter how complex, that contains any deep learning within, as a deep learning system, no matter what role deep learning might play in the larger system, even if other, more traditional elements play a critical role. To us, this seems like calling a car a transmission, just because a transmission plays an important role in the car, or a person a kidney, just because they couldn't live without at least one. Kidneys are obviously critical for human biology, but it doesn't mean that the study of medicine should be reconstrued as

281 *Ibídem*, pàg. 143. El "he" de la oració es refereix al propi Gary Marcus i a quelcom que ha escrit en un dels seus llibres en solitari *The Birth of the Mind*.

nephrology writ large. We anticipate that deep learning will play an important role in hybrid AI systems, but that doesn't mean that they will rely exclusively or even largely on deep learning. Deep learning is much more likely to be a necessary component of intelligence than to be sufficient for intelligence».[282]

En l'exemple de de Waal, el fetge i, concretament, la funció hepàtica, servia per explicar que l'estudi de l'etologia no s'interessa per un òrgan concret d'un animal en concret, sinó que això serveix per una visió global de com funciona un procés, per tant, de Waal defensava que estudiar el comportament animal que denota intel·ligència permet extrapolar coses sobre el funcionament general de la intel·ligència, i després ja es buscaria en quins òrgans aquesta es donava depenent de cada espècie i context. Aquí Marcus fa quelcom similar: parteix de la relació entre un cotxe i la transmissió per il·lustrar que, a l'igual que la transmissió només és una peça necessària pel cotxe, però no és el cotxe en si mateix, les tècniques d'aprenentatge profund poden ser una part necessària per programar la intel·ligència (ell omet el terme "programar"), però no esgoten totes les funcionalitats d'aquesta. És a dir, d'alguna forma, està fent el mateix paral·lelisme, però la diferència rau en què aquí l'òrgan a comparar és un objecte construït. Ara bé, l'analogia manté el seu paral·lelisme, cosa que descriu un comportament propi d'una etologia digital raonada. La pregunta és per què Marcus cau, aparentment de forma gens conscient, en aquest tipus de comportament, quan paràgrafs abans ha demostrat tenir una visió molt clara i tècnica del problema.

El perquè d'una etologia en Marcus i el seu impacte social

Com a mínim hi ha una frase en aquest text que pot aportar una possible resposta: l'anhel de construir una intel·ligència artificial és un anhel social, una anhel d'estar sempre acompanyats d'algú (o alguna cosa, talment una àngel de la guarda o un amic imaginari) que arribi a la mateixa conclusió, és a dir, que ens doni la raó:

P125 «Part of the reason we trust other people as much as we do is because we by and large think they will reach the same conclusions as we will, given the same evidence».[283]

Marcus creu ferament en la possibilitat i necessitat de confeccionar objectes als quals podem atorgar la mateixa confiança que a un humà. A diferència els autors de la por, és escèptic en el quan, però no en el què. Aquesta posició, que també comparteix amb matisos Mitchell i el primer Brooks, com es veurà més endavant, és segurament més perillosa que no pas els crits de Russell, Bostrom o Chalmers (cada u també de forma lleugerament diferent). Mentre que les seves exageracions poden

282 *Ibídem*, pàg. 118. Nota al peu per aclarir el concepte de "different aspects of complex problems".

283 *Ibídem*, pàg. 192.

tenir un efecte puntual en l'opinió pública, com un bon anunci publicitari, amb el temps acostumen a quedar oblidades, mentre els discursos més raonables i fonamentats poden acabar tenint una incidència més profunda a llarg termini. En altres paraules, en aquest treball es defensa que qui realment està construint i fonamentant una etologia digital amb cara i ulls és més algú com Marcus que no pas Chalmers, i que el seu comportament té molt més impacte social que no pas el discurs de la por. Per altra banda, també és cert que Marcus és algú a qui agrada la notorietat, cosa que a vegades el fa defensar posicions que poden semblar contradictòries, com quan va firmar la carta de Musk i companyia i després va dedicar cinc butlletins per puntualitzar la seva signatura. De fet, a la xarxes hi ha opinions que el veuen com una font poc fiable i tècnicament nul·la[284].

Per tant, aquest interès en ser el focus d'atenció i participar en totes les polèmiques pot penalitzar al seu nivell d'honestedat (o ingenuïtat), cosa que també afectaria al nivell d'impacte social. Per tot plegat, es podria representar la posició ambivalent de Marcus de la següent forma.

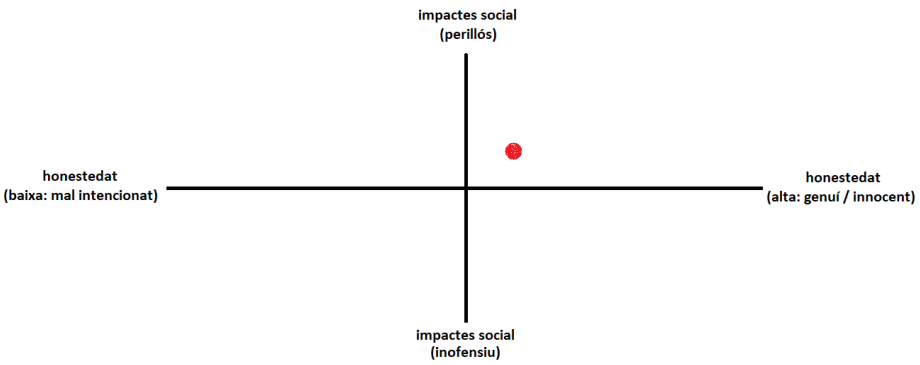

Figura 8: Nivell d'honestedat / impact social de Gary Marcus

Amb aquest gràfic s'intenta mostrar com, tot i tenir un nivell d'honestedat positiu, no puntua el màxim, i també com això condiciona en part l'impacte social, fent d'aquest discurs un perill, tot i que això també parcialment.

4.3 Melanie Mitchell i per què no hauríem de tenir por

Melanie Mitchell, una setmana més tard de la publicació de "La carta" d'Elon Musk i companyia, va escriure un text en el qual criticava aquestes iniciatives i defensava la necessitat d'una visió realista. Així replicava a un senador americà, Chris Murphy, que havia acabat de piular que la humanitat no estava preparada per la IA: P126:«Senator, I'm an AI researcher. Your

284 Wellshitsguessnot *et alii* (10.01.2024). "Gary Marcus is not an 'AI Expert'" en *Reddit*. Consultat el 16 de juliol de 2024 a: https://www.reddit.com/r/singularity/comments/1931eyy/gary_marcus_is_not_an_ai_expert/

description of ChatGPT is dangerously misinformed. Every sentence is incorrect. I hope you will learn more about how this system actually works, how it was trained, and what it's limitations are»[285]. Per Mitchell, conèixer el vertader estat de la qüestió, és a dir, què pot i què no pot fer-se utilitzant algoritmes d'IA, és clau per entendre quins són els veritables perills en contrast amb els hipotètics perills, tal i com exposa citant un article, mencionat en el capítol 3, a les autores de "Stochastic Parrots": «Those hypothetical risks are the focus of a dangerous ideology called longtermism that ignores the actual harms resulting from the deployment of AI systems today»[286]. Per Mitchell és evident que darrere d'aquesta ideologia hi ha uns interessos que res tenen a veure amb el benestar de la humanitat, i que s'agrupen sota l'etiqueta de *longtermism*.

El *llargterminisme*, aquestes autores el veuen encarnat en entitats com Future of Life Institute, Future of Humanity Institute, Centre for Effective Altruism o Centre for the Study of Existential Risk, algunes de les quals ja s'han mencionat en el capítol 2 i 3 d'aquest treball. De fet, el pare teòric d'aquest corrent no és altre que Nick Bostrom, qui amb la invenció del terme *existential risk* i defensant que no cal actuar contra el canvi climàtic (per ell és més rendible salvar 0.00000000001 per-cent dels 10^{23} habitants d'un futur llunyà que no pas mil milions dels actuals[287]), va iniciar una nova filosofia que va quallar especialment entre els anomenats super-rics. Aquesta ideologia els encoratge a defensar un suposat potencial humà basat en un domini de la naturalesa que maximitzi la productivitat econòmica, cosa que permetria una millora posthumana de l'espècie i així la colonització de l'univers. De fet, com denuncia Émile P. Torres, realment el que permet la fantasia de Bostrom és tot el contrari:

> By reducing morality to an abstract numbers game, and by declaring that what's most important is fulfilling "our potential" by becoming simulated posthumans among the stars, longtermists not only trivialize past atrocities like WWII (and the Holocaust) but give themselves a "moral excuse" to dismiss or minimize comparable atrocities in the future.[288]

285 MITCHELL, Melanie (03.04.2023). "Thoughts on a Crazy Week in AI News" en *AI: A Guide for Thinking Humans* (Melanie Mitchell substack). Consultat el 2 d'agost de 2023 a: https://aiguide.substack.com/p/thoughts-on-a-crazy-week-in-ai-news

286 GEBRU, Timnit; BENDER, Emily M.; MCMILLAN-MAJOR, Angelina; MITCHELL, Margaret (31.03.2023). "Statement from the listed authors of Stochastic Parrots on the «AI pause» letter" en *DAIR Institute*. Consultat el 2 d'agost de 2023 a: https://www.dair-institute.org/blog/letter-statement-March2023/

287 TORRES, Émilie P. (28.07.2021). "The Dangerous Ideas of «Longtermism» and «Existential Risk»" en *Current Affairs*. Consultat el 2 d'agost de 2023 a: https://www.currentaffairs.org/2021/07/the-dangerous-ideas-of-longtermism-and-existential-risk

288 *Ídem.*

Des dels altaveus que proporcionen aquestes entitats, el discurs de la por és traslladat amb consignes com la següent: «climate change poses a mere 1-in-1,000 chance of existential catastrophe, in contrast to a far greater 1-in-10 chance of catastrophe involving superintelligent machines»[289]. Per reduir les possibilitats que aquestes màquines super-intel·ligents acabin amb l'espècie humana, aquest autors proposen una inversió en més IA que permeti, en un futur llunyà, pujar (*upload*) literalment la ment de cada persona en una simulació computacional (Bostrom, al igual que Kurzweil, apel·la a «to move your mind to more durable media»[290]).

Mitchell ataca aquest tipus de discursos i, més enllà de denunciar als interessos dels emissaris de la por, també recorda que ja havia argumentat anteriorment –en resposta a un assaig narratiu ("Op-Ed") de Stuart Russell en el qual aquest presentava les idees principals del seu *Human Compatible*– que fins que no hi hagi un coneixement més profund de com funciona la intel·ligència humana (molt més entrelligada amb la resta d'habilitats socials que no pas una mera màquina de racionalitzar) mancava de sentit parlar de la possibilitat de digitalitzar una super-intel·ligència, i afegia, seguint a Hofstadter, un argument emergentista:

> P127: In other words, the intelligent part of your mind can't harness the fast-adding skills of your own neurons, and for good reason. This barrier — between the "self" that you are aware of and the detailed activity of your brain — permits the kind of thinking that matters for survival without getting overwhelmed ("addlebrained") by your own thought processes.[291]

Mitchell també feia evident el març de 2023 que textos com els de Russell o l'Op-Ed que el *Time Magazine* va donar a Eliezer Yudkowsky, camuflaven la responsabilitat humana en l'assumpte sota una narrativa basada en la suposada inexplicabilitat de «powerful digital minds»[292], és a dir, que enlloc d'assumir que són els humans qui programen, configuren i equilibren processos d'aprenentatge profund, tractaven la IA com una caixa negre autònoma l'únic que feia era generar més por vers quelcom aparentment desconegut: P128:«The "unexplainable" narrative gives rise to

289 ORD, Toby (2020). *The Precipice.* Citat en: TORRES, Émilie P. (28.07.2021). "The Dangerous Ideas of «Longtermism» and «Existential Risk»" en *Current Affairs*. Consultat el 2 d'agost de 2023 a: https://www.currentaffairs.org/2021/07/the-dangerous-ideas-of-longtermism-and-existential-risk

290 BOSTROM, Nick (2008). "Letter from Utopia" en *Nick Bostrom's Home Page*. Consultat el 2 d'agost de 2023 a: https://nickbostrom.com/utopia

291 MITCHELL, Melanie (31.10.2019). "We Shouldn't be Scared by 'Superintelligent A.I.'" en *The New York Times*. Consultat el 3 d'agost de 2023 a: https://www.nytimes.com/2019/10/31/opinion/superintelligent-artificial-intelligence.html

292 MITCHELL, Melanie (03.04.2023). "Thoughts on a Crazy Week in AI News" en *AI: A Guide for Thinking Humans* (Melanie Mitchell substack). Consultat el 2 d'agost de 2023 a: https://aiguide.substack.com/p/thoughts-on-a-crazy-week-in-ai-news

fear, and it has been argued that, to a degree, public fear of AI is actually useful for the tech companies selling it, since the flip-side of the fear is the belief that these systems are truly powerful and big companies would be foolish not to adopt them»[293]. I Mitchell defensava que, enlloc d'aturar sis mesos la investigació, el que calia era tot el contrari: més transparència per part de les empreses que monopolitzen la investigació en IA, i agrupar tota la recerca en una mena de projecte Manhattan.

Els primers indicis d'una etologia digital

Així doncs, Mitchell aposta pel coneixement: cal seguir investigant abans de fer declaracions incendiàries de dubtosa intenció. I cal ser crític amb la pròpia feina si es vol millorar. Aquest és el plantejament constant que presenta en tres dels seus recents articles: "Why AI is Harder Than We Think" (2021), "Abstraction for Deep Reinforcement Learning" (2022), "The Debate Over Understanding in AI's Large Language Models" (2023). Ara bé, entre el text de 2021 i el de 2023 es veu una progressió: si bé es parteix d'un anàlisi escèptic (2021), aquest anàlisi acaba concloent que l'únic camí viable per aconseguir intel·ligència artificial és a través de l'estudi de l'evolució i la seva imitació (2022), cosa que condueix a certa mitificació de les capacitats de les eines digitals (2023) molt similar a les observades en els autors de la por. Per tant, tot i una actitud escèptica i vigilant, el camí de Mitchell i similars condueix també a una etologia digital (i, en alguns moments, fins i tot a una mitologia digital), però aquesta ben raonada.

A continuació s'analitzen amb més detall els tres articles:

En "Why AI is Harder Than We Think", pujat a *arXiv* el 26 d'abril de 2021, Mitchell ordena històricament els motius pels quals la investigació d'IA no ha aconseguit els objectius plantejats des de la seva fundació identificant-los amb quatre fal·làcies: la fal·làcia de la continuïtat, la fal·làcia de la proporcionalitat, la fal·làcia de la metàfora mnemotècnica i la fal·làcia de l'exclusivitat del cervell.

La fal·làcia de la continuïtat, com s'ha vist anteriorment, va ser resumida amb gràcia per Stuart Dreyfus, el germà de Hubert Dreyfus: «It was like claiming that the first monkey that climbed a tree was making progress towards landing on the moon»[294]. És a dir, per molt que s'hagi pogut trobar un algoritme que simula algun aspecte de la intel·ligència, no es pot suposar que

293 *Ídem.*

294 DREYFUS, Hubert L. (2007). "Why Heideggerian AI Failed and how Fixing it would Require making it more Heideggerian" en *Artificial Intelligence,* 171 (2007), pág. 1143. Mitchell cita un document més recent: Hubert L Dreyfus. "A history of first step fallacies" en *Minds and Machines*, 22(2) (2012), pàgs., 87–99.

aquesta és com una escala, un *contínuum*, del qual només cal anar fent petites passes per aconseguir programar-la per complet.

La fal·làcia de la proporcionalitat es basa en la idea que si la programació de trets d'intel·ligència que permeten resoldre situacions aparentment difícils (com és jugar a escacs) ha estat possible, programar altres habilitats més senzilles (com el sentit comú) serà més fàcil. L'error d'aquesta hipòtesi el reconeix Marvin Minsky quan afirma: «easy things are hard»[295]. És a dir, no es pot suposar una proporcionalitat entre la dificultat que humanament ens suposa una activitat i les hores de programació que implica la seva digitalització. De fet, Mitchell cita a Hans Moravec, qui sembla justificar la dificultat en transformar digitalment un procés com el pensament humà, producte de l'evolució: «Encoded in the large, highly evolved sensory and motor portions of the human brain is a billion years of experience about the nature of the world and how to survive in it»[296]. Per tant, tot i que humanament se sigui poc conscient de com funciona la forma de pensar més instintiva, no és comparable el temps que l'espècie humana ha trigat en desenvolupar-la amb el temps que fa que s'han inventat els escacs. És una qüestió d'escala i aquesta fal·làcia obvia que la intel·ligència humana parteix d'una escala natural, no cultural.

La fal·làcia de la metàfora mnemotècnica ("The lure of wishful mnemonics" l'anomena Mitchell) denuncia com una metàfora usada primerament com a inspiració o simplement com a truc mnemotècnic es pot acabar convertint en una condemna que dirigeixi la investigació cap un rumb sense sentit, és a dir, com de desaconsellable pot ser l'ús de termes associats amb la intel·ligència humana per descriure el comportament i evolució dels programes amb IA[297]. Mitchell reconeix que la identificació d'aquesta fal·làcia es deu a Drew McDermott qui, el 1976, ja va ridiculitzar aquest ús antropomòrfic en un article titulat "Artificial intelligence meets natural stupidity", com s'ha mencionat anteriorment. La seva contraproposta era clara i així ho aconsellava a un enginyer en IA que hagués d'implementar una funció que simulés la comprensió humana: «If he calls the main loop of his program "UNDERSTAND," he is (until proven innocent) merely begging the question. What he should do instead is refer to this main loop as "G0034," and see if he can convince himself or

295 MINSKY, M. "Decentralized minds" en *Behavioral and Brain Sciences*, 3(3), 1980, pàgs. 439–440. Citat per MITCHELL, Melanie. "Why AI is Harder Than We Think" en *arXiv*. Consultat el 4 d'agost de 2023 a: https://arxiv.org/abs/2104.12871v1

296 MORAVEC,Hans (1988). *Mind Children: The Future of Robot and Human Intelligence*, Harvard University Press, 1988, pàgs. 15-16. Citat per MITCHELL, Melanie. "Why AI is Harder Than We Think" en *arXiv*, pàg. 4. Consultat el 4 d'agost de 2023 a: https://arxiv.org/abs/2104.12871v1

297 MITCHELL, Melanie (26.04.2021). "Why AI is Harder Than We Think" en *arXiv*, pàg. 5. Consultat el 4 d'agost de 2023 a: https://arxiv.org/abs/2104.12871v1

anyone else that G0034 implements some part of understanding»[298]. McDermott identificava una sèrie de conceptes que s'estaven utilitzant sense massa rigor en aquest camp, com ara, *comportament*, *resposta* o *aprenentatge*. I també, de forma implícita, descrivia el que en aquest treball anomenem la Inversió de la Metàfora Computacional: «It seems much smarter to put knowledge about translation from natural language to internal representation in the natural language processor, not in the internal representation»[299]. De fet, quan McDermott escriu aquest article, els llenguatges de programació encara són d'un nivell tan baix que un programador s'ha de preocupar de l'ordre de la pila: manquen de funcions d'alt nivell com les actuals, cosa que implica que el problema descrit per McDermott sigui avui encara més evident.

Els llenguatges de programació actual són tan similars a un pseudollenguatge que emmascaren una evidència: són llenguatges formals i no representen de la mateixa manera que el llenguatge natural (ni ho haurien de pretendre). McDermott ironitza sobre quelcom que avui dia ja no sembla broma quan proposa anomenar "conversa" a l'intercanvi d'*inputs* i *outputs* entre mòduls; i s'imagina, de nou humorísticament, la següent situació: «let the modules speak in human tongues. Let them use metaphor, allusion, hints, polite requests, pleading, flattery, bribes, and patriotic exhortations to their fellow module»[300]. Cal recordar que el 31 de juliol de 2017 va ser notícia la possibilitat que dos algoritmes d'IA haguessin inventat un llenguatge privat incomprensible pels humans[301]; tot i que la notícia va ser contextualitzada i desdramatitzada immediatament[302], denota que l'ocurrència de McDermott sembla avui un plantejament raonable. Ara bé, més enllà del to humorístic de l'article de McDermott, aquest assenyalava ja la base del problema: «The obsession with natural language seems to have caused the feeling that the human use of language is a royal road to the cognitive psyche».[303]

Aquest argument no dista molt del que s'ha citat anteriorment de Hofstadter: els conceptes mentals no han de per què coincidir amb les estructures biològiques dels quals sorgeixen. Ara bé,

298 MCDERMOTT, Drew (abril 1976). "Artificial intelligence meets natural stupidity" en *ACM SIGART Bulletin*, 57, 1976, pàgs. 4–9. Consultat el 21 d'agost de 2024 a: https://doi.org/10.1145/1045339.1045340

299 *Ibídem*, pàg. 6.

300 *Ídem*.

301 WEHNER, Mike (31.07.2017). "Facebook engineers panic, pull plug on AI after bots develop their own language" en *BGR.* Consultat el 8 d'agost de 2023 a: https://bgr.com/science/facebook-ai-shutdown-language/

302 MCKAY, Tom (31.07.2017). "No, Facebook Did Not Panic and Shut Down an AI Program That Was Getting Dangerously Smart" en *Gizmodo*. Consultat el 8 d'agost de 2023 a: https://gizmodo.com/no-facebook-did-not-panic-and-shut-down-an-ai-program-1797414922

303 MCDERMOTT, Drew (abril 1976). "Artificial intelligence meets natural stupidity" en *ACM SIGART Bulletin*, 57, 1976, pàg. 7. Consultat el 21 d'agost de 2024 a: https://doi.org/10.1145/1045339.1045340

mentre que Hofstadter assenyala la poca comprensió d'aquell moment entre cervell i ment, cosa que implica la possible comprensió en algun altre moment, McDermott es limita a assenyalar la incommensurabilitat entre l'estructura natural del parlant davant la convencionalitat del que ell anomena una gramàtica:

> Linguists have, I think, suffered from this self-misdirection for years. The standard experimental tool of modern linguistics is the eliciting of judgments of grammaticality from native speakers. Although anyone can learn how to make such judgments fairly quickly, it is plainly not a skill that has anything to do with ability to speak English. The real parser in your head is not supposed to report on its inputs' degree of grammaticality; indeed, normally it doesn't "report" at all in a way accessible to verbalization. It just tries to aid understanding of what it hears as best it can. So the grammaticality judgment task is completely artificial. It doesn't correspond to something people normally do.[304]

Mitchell recull, doncs, part d'aquesta crítica, en concret, com el vocabulari d'arrel antropomòrfica en la informàtica pot ser contraproduent, encara que sigui, en un primer moment, com abreviació (*shorthand*) o mnemotècnic, i identifica, sense aprofundir-hi, la figura conceptual que aquí anomenem Inversió de la Metàfora Computacional: P129:«Indeed, the way we talk about machine abilities influences our conceptions of how general those abilities really are»[305]. I Mitchell defensa que això pot ser enganyós (*misleading*) tant pel públic en general com pels experts en IA.

La quarta fal·làcia no fa més que aprofundir en aquesta idea: les metàfores poden portar a un reduccionisme com el de pensar que la ment o la intel·ligència és exclusivament en el cervell: «intelligence is all in the brain»[306]. Mitchell, seguint a Rodney Brooks, a qui cita explícitament, explica com el camp de la computació, des dels seus orígens, s'ha guiat per la model psicologista pel qual la ment no deixa de ser un procés d'informació: P130:«This model views the mind as a kind of computer, which inputs, stores, processes, and outputs information. The body does not play much of a role except in the input (perception) and output (behavior) stages. Under this view, cognition takes place wholly in the brain, and is, in theory, separable from the rest of the body»[307]. Ja s'ha vist com aquest supòsit condueix a que, si s'aconsegueix digitalitzar el cervell, hauria de ser possible viure eternament com a entitats digitals, cosa que Mitchell tornarà a criticar aquí. Ara bé, la

304 *Ídem*.

305 MITCHELL, Melanie (26.04.2021). "Why AI is Harder Than We Think" en *arXiv*, pàg. 5. Consultat el 4 d'agost de 2023 a: https://arxiv.org/abs/2104.12871v1

306 *Ibídem*, pàg. 6.

307 *Ídem*.

metàfora computacional és això: una metàfora, i potser una que ha arribat l'hora de desterrar, tal i com qüestiona citant a Brooks:

> Neuroscience uses computation as a metaphor, and I question whether that's the right set of metaphors […] Is information processing the right metaphor there? Or are control theory and resonance and synchronization the right metaphor? We need different metaphors at different times, rather than just computation.[308]

Mitchell cita també alguns projectes alternatius, com el Mark Johnson, que s'han centrat en la cognició encarnada (*embodied cognition*), és a dir, en entendre que la cognició es produeix a través de tot el cos, tot i que oblida que aquesta via d'investigació ja havia estat seguida pel propi Brooks (i també Walter Jackson Freeman III, com comenta Dreyfus[309]) i les seves formigues robotitzades, després de la qual ja van voler saltar a un robot humanoide, cosa que va criticar Dreyfus, més partidari d'anar pas a pas: «As in the days of GOFAI, on the basis of Brooks' success with insect-like devices, instead of trying to make, say an artificial spider, Brooks and Dennett decided to leap ahead and build a humanoid robot»[310]. Mitchell tanca l'argumentació de l'article sense deixar del tot clar si aquesta via d'investigació seria factible i, a la conclusió, es limita a defensar dues idees: P131:«It's clear that to make and assess progress in AI more effectively, we will need to develop a better vocabulary for talking about what machines can do. And more generally, we will need a better scientific understanding of intelligence as it manifests in different systems in nature»[311]. És a dir, la IA necessita més vocabulari, més eines i més coneixement de com funciona la intel·ligència, no només la humana. Defensa Mitchell que només així es podrà abordar el problema de sentit comú i poder transferir-lo a una màquina, i afegeix, en l'única nota a peu de pàgina de l'article, la següent afirmació: P132:«Some have questioned why we need machines to have *humanlike* cognition, but if we want machines to work with us in our human world, we will need them to have the same basic knowledge about the world that is the foundation of our own thinking».[312]

Fins aquesta nota a peu de pàgina, només a través de les referències fetes per Mitchell es podia haver insinuat que hi havia indicis d'una etologia digital. De fet, durant l'explicació

308 BROOKS, Rodney A. (13.05.2019). "The Cul-de-Sac of the Computational Metaphor" en *Edge*. Consultat el 30.07.2023 a: https://www.edge.org/conversation/rodney_a_brooks-the-cul-de-sac-of-the-computational-metaphor

309 DREYFUS, Hubert L. (2007). "Why Heideggerian AI Failed and how Fixing it would Require making it more Heideggerian" en *Artificial Intelligence,* 171 (2007), pàgs. 1137–1160.

310 *Ibídem*, pàg. 1142.

311 MITCHELL, Melanie (26.04.2021). "Why AI is Harder Than We Think" en *arXiv*, pàg. 8. Consultat el 4 d'agost de 2023 a: https://arxiv.org/abs/2104.12871v1

312 *Ídem*. La cursiva és de Mitchell.

d'aquestes fal·làcies no apareix cap element explícit d'etologia digital; tot el contrari, sembla denunciar, tant pel vocabulari utilitzat com explícitament per algunes de les crítiques que fa, la possibilitat d'una etologia digital. Ara bé, les referència a Moravec, Brooks i Dreyfus, i la seva explicació d'una IA més encarnada i més gradual, introdueixen una idea biològica que sí que pot implicar una etologia digital: la idea que en el camp de la IA cal seguir els passos de l'evolució (comportament 10 d'una etologia digital raonda). Aquesta idea no apareix explícitament en l'article (de fet, l'arrel *evolv-* només hi apareix una vegada i en la citació de la paradoxa de Moravec), però sí en Dreyfus, a qui Mitchell cita fent referència a textos en els quals la idea evolutiva hi és present; per exemple, la proposta d'una IA connexionista de Dreyfus entén l'aprenentatge reforçat (*reinforcement learning*) com una manera en què una màquina aprengués de l'experiència:

> The meaning of the input is neither in the stimulus nor in a mechanical response directly triggered by the stimulus. Significance is not stored as a memory-representation nor an association. Rather the memory of significance is in the repertoire of attractors as classifications of possible responses – the attractors themselves being the product of past experience.[313]

L'acceleració d'aquesta experiència és la que permetria simular una forma d'evolució i, de fet, així s'anomenen aquestes tècniques de programació (*Evolution Strategies*) i així les defineixen:

> At every iteration ("generation"), a population of parameter vectors ("genotypes") is perturbed ("mutated") and their objective function value ("fitness") is evaluated. The highest scoring parameter vectors are then recombined to form the population for the next generation, and this procedure is iterated until the objective is fully optimized.[314]

Per tant, la teoria d'una IA encarnada sembla passar per una conversió de les teories de la computació a les teories de l'adaptació, precisament un dels pilars que trobava a faltar Brooks: «The way we engineer our computational systems is with no adaptation, and the way all biological systems work is through adaptation at every level all the time».[315]

A banda d'aquestes referències a les quals Mitchell cita, però no acaba d'abraçar clarament, és més rellevant, com s'ha mencionat abans, la mateixa nota a peu de pàgina: P132:«Some have questioned why we need machines to have *humanlike* cognition, but if we want machines to work

313 DREYFUS, Hubert L. (2007). "Why Heideggerian AI Failed and how Fixing it would Require making it more Heideggerian" en *Artificial Intelligence,* 171 (2007), pàg. 1162.

314 SALIMANS, Tim; HO, Jonathan; CHEN, Xi; SIDOR, Szymon; SUTSKEVER, Ilya (2017). "Evolution strategies as a scalable alternative to reinforcement learning" en *arXiv*, 2017, pàg. 2. Consultat el 21 d'agost de 2024 a: arxiv.org/pdf/1703.03864

315 BROOKS, Rodney A. (13.05.2019). "The Cul-de-Sac of the Computational Metaphor" en *Edge*. Consultat el 30.07.2023 a: https://www.edge.org/conversation/rodney_a_brooks-the-cul-de-sac-of-the-computational-metaphor

with us in our human world, we will need them to have the same basic knowledge about the world that is the foundation of our own thinking»[316]. L'explicació és tan poc convincent que cal que estigui en una nota a peu de pàgina: sembla fruit d'un comentari fet en procés de revisió (*some have questioned*) i la seva justificació requeriria d'un apartat sencer, cosa que no sembla encaixar amb l'estructura de l'article (no s'hauria d'afegir informació a la conclusió). Tanmateix, Mitchell ha de ser conscient que això no és tan obvi com ella defensa. De fet, el mateix McDermott, a l'article citat per Mitchell, planteja indirectament just això que ella exclou: «Perhaps we should postpone trying to get computers to speak English, and try programming librarians in PL/1»[317]. És a dir, enlloc de centrar els esforços en desenvolupar unes eines que imitin el comportament humà fins el punt d'utilitzar un llenguatge natural amb destresa, tal vegada seria més convenient limitar-se a utilitzar llenguatges clarament formals que permetin centrar-se en la feina que es vol resoldre amb aquell algoritme en qüestió. També Brooks sembla que és cada vegada més partidari de fer eines que, tot i interactuar amb els humans, per exemple, en el seu lloc de treball, restin clarament sota el seu control, com es veurà més endavant: «But then the magic of our robot is that it looks like a shopping cart. It's got handlebars on it. If a person goes up and grabs it, it's now a powered shopping cart or powered cart that they can move around. So [the warehouse workers] are not subject to the whims of the automation»[318].

Per tant, d'alguna forma o altra Mitchell ha de ser conscient que existeix aquesta alternativa, és a dir, una proposta que mantingui l'eina com a eina i la IA al servei humà i centrada en les necessitats humanes. I també Mitchell ha de ser conscient, i de fet, ho és, que l'alternativa és una etologia digital, és a dir, algun tipus de pretensió de creació divina. Mitchell acaba l'article recordant com, tant Terry Winograd l'any 1977 com Eric Horvitz, el 2017 s'havien referit a la IA com la nova alquímia, i reclama un canvi: P133:«In order to understand the nature of true progress in AI, and in particular, why it is harder than we think, we need to move from alchemy to developing a scientific understanding of intelligence»[319].

316 MITCHELL, Melanie (26.04.2021). "Why AI is Harder Than We Think" en *arXiv*, pàg. 8. Consultat el 4 d'agost de 2023 a: https://arxiv.org/abs/2104.12871v1. La cursiva és de Mitchell.

317 MCDERMOTT, Drew (abril 1976). "Artificial intelligence meets natural stupidity" en *ACM SIGART Bulletin*, 57, 1976, pàg. 7. Consultat el 21 d'agost de 2024 a: https://doi.org/10.1145/1045339.1045340

318 ZORPETTE, Glenn (17.05.2023). "Just Calm Down About GPT-4 Already And stop confusing performance with competence, says Rodney Brooks" en *IEEE Spectrum*. Consultat el 8 d'agost de 2023 a: https://spectrum.ieee.org/gpt-4-calm-down

319 MITCHELL, Melanie (26.04.2021). "Why AI is Harder Than We Think" en *arXiv*, pàg. 8. Consultat el 4 d'agost de 2023 a: https://arxiv.org/abs/2104.12871v1

Ara bé, si només s'analitza aquest article seria complicat defensar que Mitchell col·labora amb una etologia digital: no s'hauria de poder caracteritzar una forma de pensar per una mera nota a peu de pàgina i la referència a altres autors. Tanmateix, aquestes idees tornaran a aparèixer i cada vegada de forma més recurrent en els següents dos articles.

Una proposta igualment etològica

En "Abstraction for Deep Reinforcement Learning", pujat a *arXiv* el 29 d'abril de 2022, Mitchell, juntament amb Murray Shanahan, analitzen els desenvolupaments fets en IA per veure quins d'aquests poden millorar la capacitat d'abstracció en el context de l'aprenentatge profund per reforç. Com en l'anterior article, el text no contribueix de forma malintencionada a una etologia digital, ja sigui confonent el lector amb metàfores mal explicades o usos poc acurats dels termes, sinó que parteix de la idea que cal fer una etologia digital per aconseguir una IA útil. Per altra banda, a diferència de l'article anterior, en aquest cas aquesta idea no és merament implícita o atribuïda a altres, sinó que s'abraça i es defensa de forma explícita.

La primera línia de l'article ja deixa clar que el subjecte d'aplicació no són només els humans: P134:«Consider an embodied agent, either an animal (human or nonhuman), a physical robot, or a virtual agent in a simulated environment»[320]. La primera proposició del text és un axioma etològic: una agent encarnat[321] és tota aquella entitat dotada de cos, ja sigui físic o virtual. Per tant, sense cap subterfugi estableixen que es pot estudiar les entitats digitals i els robots com si fossin qualsevol altre animal. Aquesta segurament és la màxima expressió (i la més honesta) per confeccionar una etologia digital: declarar qualsevol entitat virtual com equivalent a qualsevol animal en la mesura que tots tenen cos (encara que sigui només virtual). I és una conseqüència d'haver assumit, com feia Mitchell en l'article anterior, que per aconseguir arribar a una veritable intel·ligència artificial s'havien de seguir els passos de l'evolució: primer hi havia cossos, d'alguns d'ells en va sorgir intel·ligència, i d'alguns d'aquests, intel·ligència humana, que és l'objectiu final a digitalitzar.

320 SHANAHAN, Murray; MITCHELL, Melanie (29.04.2022). "Abstraction for Deep Reinforcement Learning" en *arXiv*, pàg. 1. Consultat el 4 d'agost de 2023 a: https://arxiv.org/abs/2104.12871v1

321 Una altra possible traducció seria *incorporat*, com ho fan a l'italià amb *cognizione incorporata*, però té el problema que *incorporat* és un adjectiu que es fa servir en un altre sentit en català (a saber, que portes sempre a sobre); o també *personificat*, com quan clarament substitueix a una persona (el problema aquí és que no sempre substitueix una persona); aquí s'ha optat per l'estil francès que tradueix el concepte cognitiu per *cognition incarnée*, tot i que el concepte de *cos* no inclou, especialment en aquest context, el de *carn*. Altres opcions serien *agent cossat* («Que té el tronc configurat de tal o tal manera» segons la primera accepció de *cossat* del DIEC) o algun llatinisme com *agent in corpore* o *ad corpus*, però si s'optava per no utilitzar el català, es podia haver deixat l'opció anglesa original.

Mitchell i Shanahan es plantegen què deu pensar un animal abans d'actuar («How can I shape my future experience to my liking?»[322]) i assumeixen, sense cap explicació, que això es tradueix en una certa forma de tenir en compte les experiències passades («How does my present experience resemble what I have experienced in the past?»[323]). Feta aquesta reducció, assimilen el problema de la corporeïtat (o encarnació) a un problema de la generalització: P135:«How well an agent does this — that is to say how well it generalises from past experience — is one measure of its intelligence».[324]

Aquest enfocament planteja alguns problemes que s'analitzen per ordre d'aparició: abans de res, la major part d'actuacions animals són instintives i, com a tals, ometen la pregunta de com modelar les experiències futures al gust de cada u. La discussió en el camp de l'etologia (i aquí sí que la resposta s'ha de donar respectant aquest camp si es pretén cenyir-se a l'estudi dels anomenats agents encarnats) és com actuen els espècimens tenint en compte els instints i com s'adapten a un entorn concret amb més o menys encert, és a dir, com hi reaccionen, cosa molt diferent de com es plantegen el futur. Una de les respostes habituals és que hi ha un influència mútua entre les actuacions filogenètiques i les apreses, tal i com defensa la Teoria dels Sistemes de Desenvolupament (DST per la seva abreviació en anglès). Per tant, sembla que la pregunta premeditada de com enfocar el futur només tindria sentit en accions molt concretes d'alguns individus molt concrets, si més no amb aquests termes de consciència.

La segona pregunta que es plantegen Mitchell i Shanahan pretén cenyir la primera en un encaix massa petit: si tota experiència futura s'ha de basar en el passat, no hi ha oportunitat d'experiència nova, crítica que ja havia plantejat Lovelace en la Nota G[325] i que Turing incorpora, com a sisena objecció, en el seu article de 1950, "Computing Machinery and Intelligence"[326]. Qualsevol animal improvisa, no es limita a reproduir l'experiència passada, per tant, caldria analitzar què vol dir improvisar i, en quina mesura, no consisteix, per definició, en fer quelcom imprevisible; així que, si aquesta improvisació es basés en accions preestablertes, deixaria de ser imprevisible i, si ho fes en una funció aleatòria, no simularia la imprevisibilitat, ja que mancaria de sentit, és a dir, orientació.

322 SHANAHAN, Murray; MITCHELL, Melanie (29.04.2022). "Abstraction for Deep Reinforcement Learning" en *arXiv*, pàg. 1. Consultat el 4 d'agost de 2023 a: https://arxiv.org/abs/2104.12871v1

323 *Ídem.*

324 *Ídem.*

325 LOVELACE, Augusta Ada (1842). "Nota G" en *On Sketch of the Analytical Engine Invented by Charles Babbage*. Consultat el 17 d'agost de 2023 a: https://notage.org/

326 TURING, Allan M. (1950). "Computing Machinery and Intelligence", en *Mind*, núm. 49, pàgs. 433-460 (trad. cast.: ¿Puede pensar una máquina?, Oviedo, KRK Ediciones, 2012).

En tercer lloc, hi ha el problema de reduir la corporeïtat al problema de la generalització: tan cert és que es pot generalitzar sense cos (procés clàssic d'inducció) com que els humans fem moltes altres accions amb el cos (tocar, olorar, moure'ns), accions que després poden relacionar-se amb la presa de decisions, fins i tot, la meditada, com, per exemple, imaginar (tal i com es planteja Salvador Moya: els primers homínids en crear un ganivet, primer van haver d'imaginar-lo[327], i no l'haguessin pogut imaginar si no s'haguessin tallat diverses vegades). La facultat d'imaginar s'havia menystingut en filosofia durant molt segles, entenent-la merament com la capacitat de captar i crear imatges, fins que Hume i Kant la recuperen i li donen valor: Hume la responsabilitza de la capacitat d'associar idees i Kant de la confecció dels esquemes, pont entre la sensibilitat i l'enteniment. En el camp de la IA també se sol minimitzar el paper de la imaginació i es limita a la possibilitat de reconèixer imatges, mentre que se sol identificar amb la creació a la capacitat de confeccionar-les. Rarament se sol fer un esforç per analitzar adequadament el concepte, tret d'algunes excepcions més centrades en el món artístic, com la de Claudo Celis i María Jesús Schultz, que intenten fer un pont entre els LLM entesos com mecanismes d'associació d'idees i l'obtenció de patrons en imatges com un mecanisme d'esquematització[328] (de fet, citen a Dan McQuillan, qui enllaça directament la ciència de dades amb el neoplatonisme: «Data science can be understood as an echo of the neo-platonism that informed early modern science in the work of Copernicus and Galileo. That is, it resonates with a belief in a hidden mathematical order that is ontologically superior to the one available to our everyday senses»[329]).

Resumint, hi ha tres conceptes que Mitchell i Shanahan ignoren, com són l'instint, la improvisació i la imaginació, i posen tot el pes en la generalització (que assumeixen que consisteix en algun tipus d'abstracció en el següent paràgraf, sense contemplar la diferència entre generalitzar i abstraure, conceptes diferenciats tant conceptualment com específicament en l'àmbit de la programació[330]). Per altra banda, si tota actuació requerís una presa de decisió en el sentit que descriuen, hi hauria una regressió al passat que hauria d'acabar topant amb un principi innat d'algun

327 CORBELLA, Josep; CARBONELL, Eudald; MOYÀ, Salvador; SALA, Robert (2000). *Sapiens: El llarg camí dels homínids cap a la intel·ligència*, Proteus, Barcelona, 2000, pàg. 31.

328 CELIS, Claudio; SCHULTZ, María Jesús (2021). "Notes on an Algorithmic Faculty of the Imagination" en *Anthropocenes – Human, Inhuman, Posthuman*, 2(1): 12, 2021. Consultat el 17 d'agost de 2023 a: https://www.anthropocenes.net/article/1016/galley/4928/view/.

329 MCQUILLAN, Dan (21.08.2017). "Data Science as Machinic Neoplatonism" en *Philosophy & Technology*, 31 pàgs. 253–272, 2018. Consultat el 18 d'agost de 2023 a: https://doi.org/10.1007/s13347-017-0273-3

330 AABY, Anthony A. (1996). "Abstraction and Generalization" en *Internet Archive Wayback Machine*. Consultat, el 18 d'agost de 2023 a: https://web.archive.org/web/20180328151725/http://www.emu.edu.tr:80/aelci/courses/d-318/d-318-files/plbook/absgen.htm

tipus, que Mitchell i Shanahan tampoc es plantegen, però que ja va ser plantejat des del punt de vista simbòlic per Roger Shank entre d'altres el 1969[331]. Per contra, es plantegen el procés *ad futurum*: sigui x una experiència qualsevol del present, cal trobar quines característiques té comunes amb alguna experiència del passat (se suposa que de resultat considerat òptim) per preveure una acció futura. Ara bé, quina experiència del passat se selecciona entre totes les possibles experiències tampoc és analitzat a l'article i és el conegut com el problema de l'enquadrament (*the frame problem*[332]).

Aquest problema clàssic ha estat abordat des de diferents vessants (tant la lògica com altres més epistemològiques), des de les seves primeres formulacions per McCarthy i Hayes (lògiques), en l'article titulat "Some Philosophical Problems from the Standpoint of Artificial Intelligence"[333], així com també per Dreyfus en una proposta anomenada connexionista (que pretenia abordar pròpiament els problemes epistemològics), caracteritzada per Haugeland com arrelada i encarnada (*embedded and embodied*[334]). Alguns autors diuen que, en la mesura que els LLM han estat capaços d'obtenir resultats excel·lents tant en reconeixement d'imatges com en traducció i resposta automàtica, implica que el problema de l'enquadrament ha quedat resolt, cosa que és certa a nivell lògic, però que és més que dubtosa a nivell epistemològic si es té en compte, com la pròpia Mitchell ha denunciat diferents vegades (entre d'elles, en l'article analitzat anteriorment), que els patrons que permeten fer les prediccions encertades als LLM no solen assemblar-se als patrons que els humans veiem, ja que es basen en dreceres d'aprenentatge (*shortcut learning*): P136:«learning statistical associations in the training data that allow the machine to produce correct answers but sometimes for the wrong reasons»[335]. En qualsevol cas, sembla que Mitchell i Shanahan han d'assumir una sèrie d'implícits per poder tractar el tema de l'abstracció, que és del que va realment l'article.

---

331 Citat en NILSSON, Nils J. (2010). *The Quest for Artificial Intelligence. A History of Ideas and Achievements*, New York, Cambridge University Press, 2010, pàg. 155.

332 "The frame problem". *Stanford Encyclopedia of Philosophy*. Consultat el 17 d'agost de 2023 a: https://plato.stanford.edu/entries/frame-problem/

333 MCCARTHY, John; HAYES, Patrick J. (1969). "Some Philosophical Problems from the Standpoint of Artificial Intelligence" en *Machine Intelligence 4,* B. Meltzer & Donald Michie (eds.), Edinburgh University Press., 1969, pàgs. 463--502. Consultat el 18 d'agost de 2023 a: https://www-formal.stanford.edu/jmc/mcchay69.pdf. La patent del nom ha estat qüestionada per Minsky, qui argumenta que fou ell qui va popularitzar el tractament a través de marcs temàtics (*frames*) d'aquest problema.

334 HAUGELAND, John (1993). "Mind embodied and embedded" en *Having Thought. Essays in the Metaphysics of Mind,* Cambridge (USA), Harvard University Press, 1998, pàgs. 207.

335 MITCHELL, Melanie (26.04.2021). "Why AI is Harder Than We Think" en *arXiv*, pàg. 3. Consultat el 18 d'agost de 2023 a: https://arxiv.org/abs/2104.12871v1

Després de justificar que s'opta per centrar-se en l'aprenentatge profund per reforç degut als avanços de l'última dècada, Mitchell i Shanahan fan una revisió de les diferents aproximacions al tractament de l'abstracció, la seva possibilitats d'aplicar-se, els seus progressos i també els reptes que hi resten. El primer pas és definir: P137:«We will use the umbrella term *abstraction* to denote this cluster of operations: seeing similarity (analogy-making), forming a concept, and applying a concept»[336]. Aquesta definició planteja un problema evident: no es pot trobar cap similitud si prèviament no hi ha quelcom a comparar, per tant, com s'havia ja apuntat en el paràgraf anterior, caldrà garantir un joc d'idees innates mínimes a un agent abans de reclamar-li la primera abstracció. Aquest conjunt d'idees innates mínimes, Mitchell i Shanahan l'anomenen sentit comú:

> P138: Humans and, to a lesser degree, other animals possess a repertoire of these fundamental concepts that includes, in the domain of everyday physics, such concepts as object, path, obstacle, portal, container, and so on, and, in the social domain, such concepts as other agents, being with others, meeting, giving, taking, helping, and so on.[337]

Més enllà de si el sentit comú és el menys comú de tots els sentits, i en quin sentit ho és, sembla que la definició inclou tant elements de caràcter sensomotriu (com el domini de la física diària), com d'altres de caràcter més conceptual (com el domini social). No sembla que sigui del mateix ordre la capacitat d'evitar xocar amb altres cossos i la de saludar al creuar-se amb algú i, tenint en compte que es torna a apel·lar explícitament a l'àmbit de la biologia, potser caldria tenir en compte que la primera capacitat la domina qualsevol ésser viu, mentre que la segona només els éssers vius socials. En qualsevol cas, és una mostra més de que la intenció és afegir els objectes digitals en una etologia, és a dir, fer una etologia digital. De fet, això s'acaba de fer explícit quan Mitchell i Shanahan citen l'article de Crosby *et. al*. (entre aquest *alii* hi ha Shanahan) titulat "The Animal-AI Testbed and Competition" (2020), que tal i com expliquen els seus autors, va ser un competició entre agents digitals basada en un conjunt de proves inspirades en la literatura sobre cognició animal. Només amb quatre frases de l'*abstract* queda clar que el projecte és una etologia digital:

> Modern machine learning systems are still lacking in the kind of general intelligence and common sense reasoning found, not only in humans, but across the animal kingdom. Many animals are capable of solving seemingly simple tasks such as inferring object location through object persistence and spatial elimination, and navigating efficiently in out-of-distribution novel environments. Such tasks are difficult for AI, but provide a natural stepping stone towards the

336 SHANAHAN, Murray; MITCHELL, Melanie (29.04.2022). "Abstraction for Deep Reinforcement Learning" en *arXiv*, pàg. 1. Consultat el 18 d'agost de 2023 a: https://arxiv.org/abs/2104.12871v1

337 *Ibídem*, pàgs. 1-2.

> goal of more complex human-like general intelligence [...]. In this paper we outline the environment, the testbed, the results of the competition, and discuss the open challenges for building and testing artificial agents capable of the kind of nonverbal common sense reasoning found in many non-human animals.[338]

Per Mitchell i Shanahan aquestes proves tenen un problema, que resumeixen de la següent forma: mentre que els infants d'entre 6 i 10 anys poden superar els testos que agents digitals són incapaços de fer, aquests infants es beneficien de dues coses: P139:«several years' experience observing and interacting with the real world in all its variety and complexity, not to mention the innate endowment of human evolutionary history»[339]. Cap d'aquests dos elements sembla digitalitzable: el primer perquè la vida d'aquests infants ha estat en societat; el segon, com a mínim per manca de temps.

En aquest sentit, els límits d'una etologia digital es fan evidents: la creació humana d'objectes digitals no compta amb la mateixa disponibilitat de recursos que la generació natural de subjectes naturals, ni en quant a temps ni en quant a materials. Sembla que la mera intenció de buscar una equivalència, pot resultar lleugerament pretensiós i pot portar a cometre fal·làcies com la següent: reduir la similitud al reconeixement de patrons. Mentre que les similituds es reconeixen en un entorn continu, els patrons s'extreuen d'un context discret, com assenyala Brian Cantwell Smith en la citació que Mitchell i Shanahan incorporen. De fet, la diferència entre la continuïtat de la naturalesa i la discreció de la digitalització rau en què en la primera no hi ha un lloc fixe per on tallar i les similituds tant es poden trobar a un nivell com en un altre (ja sigui visual, auditiu, olfactiu, tàctil o la barreja de tots ells), mentre que la discreció ho és en la mesura que està perfectament definit el salt entre un valor i altre, així com el nombre i possibilitats de barreges.

Per tant, sembla que Mitchell i Shanahan són conscients de les dificultats de construir, amb sentit, una etologia digital, tot i que això no els impedeix de seguir intentant-ho, per exemple, quan analitzen com retallar el temps d'aprenentatge a través d'un procés d'exposició ràpida al llenguatge: P140:«exposure to human language in infancy can be thought of as turbocharging the process of acquiring low-level abstractions from sensorimotor interaction»[340]. Tanmateix, conclouen que fins

338 CROSBY , Mathew; BEYRET, Benjamin; SHANAHAN, Murray; HERNÁNDEZ-ORALLO, José; CHEKE, Lucy; HALINA, Marta (2020). "The Animal -AI Testbed an Competition" en *Proceedings of Machine Learning Research*, 123, 2020, pàg. 164. Cal assenyalar que l'article va ser finançat pel Leverhulme Centre for the Future of Intelligence, entitat que ja havia aparegut també en capítols anteriors.

339 SHANAHAN, Murray; MITCHELL, Melanie (29.04.2022). "Abstraction for Deep Reinforcement Learning" en *arXiv*, pàg. 2. Consultat el 18 d'agost de 2023 a: https://arxiv.org/abs/2104.12871v1

340 *Ibídem*, pàg. 6.

que no hi hagi els corpus de dades suficientment grans i amb les propietats estadístiques adequades (per evitar biaixos), ni els LLM com el GPT poden anar massa més lluny en la possibilitat d'abstracció. Ara bé, resulta curiós que el terme utilitzat per descriure què pot suposar l'exposició al llenguatge sigui *turbocarregador* (enlloc de l'habitual *turbocompressor*), és a dir, que s'hagi de recórrer a un concepte mecànic per explicar un procés natural (és dubtós que els autors tinguin en ment l'etimologia de l'arrel *turba-*: massa en moviment circular[341]), cosa que implicaria que, tot i que aquest projecte d'etologia digital s'ha construït de forma explícita contra l'ús de metàfores mal explicades o usos poc acurats dels termes, al final hagi d'acabar recorrent a aquestes metàfores mateixes per explicar-se. En qualsevol cas, Mitchell i Shanahan conclouen que el problema té difícil solució mentre no se sàpiga lligar qualsevol variable i ranura amb qualsevol altra, és a dir, que és una forma de dir que el problema és la discreció mateixa.

Per tant, com s'ha vist, en aquest article Mitchell (i Shanahan) abracen explícitament la literatura d'arrel biològica i analitzen quin encaix pot tenir en el món tecnològic digital, és a dir, construeixen una etologia digital sense complexos. Aquest estil desacomplexat la portarà, el següent any, a presentar un article amb algunes característiques similars a les que utilitzen els autors de la por.

D'una etologia digital a una mitologia digital?

En "The Debate Over Understanding in AI's Large Language Models", pujat a *arXiv* el 10 de febrer de 2023, Mitchell, juntament amb David C. Krakauer (president del Santa Fe Institute, organisme on treballa Mitchell), exposa els arguments a favor i en contra sobre si es pot afirmar que els LLM entenen el llenguatge natural[342]. El motiu per fer aquesta exposició és, segons ells, que el debat no és merament acadèmic: P141:«the extent and manner in which machines understand our world has real stakes for how much we can trust them to drive cars, diagnose diseases, care for the elderly, educate children, and more generally act robustly and transparently in tasks that impact humans».[343]

Aquest argument, o part de l'argument, ja l'havia fet servir Mitchell en l'article de 2021, quan havia afirmat, sense més explicació, que una eina que interactua amb els humans ha d'entendre'ls – encara que no se li exigeixi el mateix a un avió, per exemple, tot i que interactua també amb els humans i, al mateix temps, té un elevat grau d'autonomia en molts moments del vol; se li exigeix

341 Turborreactor. *Diccionario Etimológico Castellano En Línea* (DECEL). Consultat el 19 d'agost de 2023 a: https://etimologias.dechile.net/?turborreactor

342 MICTHELL, Melanie; KRAKAUER, David C. (10.02.2023). "The Debate Over Understanding in AI's Large Language Models" en *arXiv*, pàg. 1. Consultat el 22 d'agost de 2023 a: https://arxiv.org/abs/2210.13966

343 *Ídem.*

als enginyers aeronàutics que l'han dissenyat, no a l'avió, que l'avió no s'estavelli sol i que funcioni tal i com està previst. Aquí s'hi afegeix un element nou: aquesta interacció té a veure, d'alguna manera, amb la capacitat de comprendre i comunicar. Per tant, Mitchell i Krakauer justifiquen l'interès en el tema amb la següent relació: com que els humans utilitzem el llenguatge natural per descriure i comprendre el món físic i social en el qual ens movem i els LLM utilitzen llenguatge natural per respondre les preguntes que els humans els fem, llavors és pensable que les seves respostes impliquin que comprenguin aquest mateix entorn físic i social. Sembla que Mitchell i Krakauer troben suficient aquesta relació, sense preguntar-se si no es podria haver plantejat aquest mateix argument abans de l'invent dels LLM, quan l'idioma d'intercanvi d'instruccions era un llenguatge formal, que també els humans fem servir per descriure i entendre el món físic i social que ens envolta.

Com s'ha vist anteriorment, McDermott ja havia assenyalat que el debat sobre la comprensió (*understanding*) s'acabaria de cop si els investigadors es limitessin a programar funcions amb noms sense significat aparent com G0034 enlloc d'UNDERSTAND. Aquesta idea, tot i l'aspecte humorístic a través del qual la presentava, ja assenyalava el problema de fons, és a dir, la suposició que hi ha una relació natural entre les idees, les paraules i les coses, i que l'aparent capacitat d'articular un dels dominis permetia, per transitivitat, suposar la capacitat d'articular en els altres dominis. Mitchell, i Krakauer ben segur que també, coneixen la crítica de McDermott, però tanmateix, enlloc d'assenyalar les fal·làcies del mateix enunciat de la qüestió, opten per el plantejament següent: P142:«Moreover, the current debate suggests a fascinating divergence in how to think about understanding in **intelligent systems**, in particular the contrast between **mental models** that rely on statistical correlations and those that rely on causal mechanisms»[344]. Tot i que ho dissimularan sota una cadena de preguntes retòriques, aquesta serà la tesi que defensaran en aquest article: es pot admetre que els LLM tenen comprensió si s'accepta que són sistemes intel·ligents, que les correlacions estadístiques generen coneixement i que aquest coneixement és traslladat a través d'un llenguatge natural. A aquesta conclusió, plantejada com a dilema a l'inici, s'hi arribarà després de resseguir els diferents posicionaments respecte la comprensió, des dels més favorables a afirmar, no només que és factible la intel·ligència artificial, sinó que ja s'està donant, fins els més escèptics (*AI denialism* ho anomenen, expressió que actualment assumeix una certa càrrega pejorativa a l'associar-se als moviments negacionistes com els del clima, les vacunes o altres evidències científiques), passant per alguns punts entremig.

344 *Ídem.* La negreta és pròpia i serveix per marcar el terme etològic en discussió.

En qualsevol cas, es constata que el plantejament del debat que fan Mitchell i Krakauer parteix de la idea que un sistema intel·ligent (i aquí inclouen algoritmes en aquesta denominació) pugui tenir un model mental, cosa que implica tenir una ment. Per tant, plantegen un debat esbiaixat des del qual ja serà impossible defensar si un sistema intel·ligent té o no un model mental, ja que es dona per fet que allò a debatre és entre diferents models mentals. En aquest sentit, la descripció del debat en aquests termes és, en ella mateixa, etològica i impossibilita la pregunta pel propi sentit d'una etologia digital i, conseqüentment, la seva negació.

Segons Mitchell i Krakauer, abans de l'èxit recent dels LLM (situat entre 2022 i inici de 2023 presumiblement), hi havia un acord general dins la comunitat investigadora en IA: P143:«while AI systems exhibit seemingly intelligent behavior in many specific tasks, they do not *understand* the data they process in the way humans do»[345]. És a dir, no és que no la comprenguessin, sinó que no la comprenia de la mateixa manera com ho fan els humans, que és diferent que simplement no comprendre-la. Afegeixen que aquest acord s'ha trencat amb la popularització dels LLM dels últims anys, cosa que ha portat a un sector de la comunitat investigadora a fer afirmacions com les següents: P144:«A threshold was reached, as if a **space alien** suddenly appeared that could communicate with us in an eerily human way. Only one thing is clear—**LLMs are not human…** Some aspects of their behavior appear to be **intelligent**, but if not human intelligence, what is the nature of their intelligence?».[346]

Aquest plantejament, que recorda als que s'han analitzat quan s'han estudiat els denominats autors de la por, no només assenta que els LLM no són humans (i pel sentit del text, això no vol dir en cap cas que els consideri inferiors als humans), sinó que accepta també que tenen intel·ligència i que aquesta supera les mesures humanes (la comparació amb un alienígena ja s'ha vist anteriorment que serveix per sobredimensionar el fenomen). De fet, sense utilitzar paraules d'altres, Mitchell i Krakauer ja fan alguna afirmació que també s'ha vist en autors com Russell i similars: P145:«How LLMs perform these feats remains **mysterious** for lay people and scientists alike»[347] o P146:«The neuroscientist Terrence Sejnowski described the **emergence** of LLMs this way»[348]. Com s'ha vist anteriorment, rodejar de misteri i dotar d'autonomia la fabricació d'un conjunt de línies de codi constitueix dos trets bàsics de la confecció d'una etologia digital intuïtiva. Ara bé, quan

345 *Ídem.* La cursiva és dels autors de l'article.

346 SEJNOWSKY, Terrence. Citat per MICTHELL, Melanie; KRAKAUER, David C. (10.02.2023). "The Debate Over Understanding in AI's Large Language Models" en *arXiv*, pàg. 2. Consultat el 23 d'agost de 2023 a: https://arxiv.org/abs/2210.13966. La negreta és pròpia i serveix per marcar el terme etològic en discussió.

347 *Ídem*.

348 *Ídem*.

aprofundeixen una mica més en els arguments dels negadors d'IA, admeten que aquests afirmen que models com el GPT o LaMDA no poden tenir comprensió perquè manquen d'experiència o de models mentals del món: P147:«their training in predicting words in vast collections of text has taught them the form of language but not the meaning»[349]. L'argument, tot i que pretén reforçar-se amb la citació d'alguns autors alineats amb aquest (citen Bender, Gebru, LeCun, Gopnik o Marcus), l'acaben concloent de la següent forma, que no sembla seguir una deducció lògica evident: P148:«Those on the "LLMs do not understand" side of the debate argue that while the fluency of large language models is surprising, our surprise reflects our lack of intuition of what statistical correlations can produce at the scales of these models».[350]

Així doncs, el problema de fons no és que els LLM tinguin o no tinguin comprensió (problema plantejat pels *AI denials*), sinó la manca d'intuïció dels humans sobre les correlacions estadístiques, cosa que també col·labora en la confecció d'un plantejament etològic. Per una banda, a través d'una confusió terminològica: en un sentit tècnic, associar la intuïció (tradicionalment, la intuïció, a diferència del raonament discursiu, consisteix en la captació immediata d'una idea, primera fase de la intel·lecció, per tant, part de la facultat de l'enteniment) amb la troballa de patrons estadístics és, valgui la redundància, contraintuïtiu: la troballa de patrons entre milers de dades no és una tasca intuïtiva (si més no, humanament), sinó deductiva, és a dir, cal aplicar diferents passos analítics fins a trobar-la; tampoc ho és per un procés que, com indica els seu nom, ha de seguir una sèrie de passos que han estat programats, encara que sigui en un llenguatge formal, de forma discursiva (en aquest cas, lògica), és a dir, línia rere línia[351]. Així, associar la intuïció amb les correlacions estadístiques és, com a mínim, contraintuïtiu i genera una estranya confusió que permet barrejar conceptes *de iure* separats. Per altra banda, també és una forma de minorització de les capacitats humanes (aspecte necessari si es pretén sobredimensionar les capacitats dels LLM), ja que els humans triguem hores en trobar aquestes correlacions (el procés informàtic sembla immediat, però internament passa per milers de línies de codi, cosa que també manca d'immediatesa en sentit estricte: només des del punt de vista de l'observador extern és tan ràpid que sembla immediat). Per tant, tot i que Mitchell i Krakauer pretenen fer una descripció objectiva de

349 *Ibídem*, pàg. 3.

350 *Ídem*.

351 Contraargumentar dient que qualsevol intuïció humana també s'ha de generar passant per milers de neurones, cosa que implicaria que internament no és immediata, no té en compte que una cosa és el codi informàtic, per molt connexionista que aquest sigui, i el codi genètic i les connexions sinàptiques que s'hi regulin. I si ho té en compte, llavors estarà assumint la pretensió que el codi informàtic és un intent d'emulació del codi genètic (ignorant la part química del seu funcionament, com a mínim) i, per tant, construint una etologia digital, que és el que es volia demostrar.

l'estat de la qüestió, hi ha una sèrie de característiques en el seu plantejament i, en concret, la manera en com fan el plantejament i la tria de termes per fer-lo, que porten cap a la conclusió mencionada: els LLM tenen comprensió si acceptem que existeix una forma de comprensió estadística.

Abans de concloure, Mitchell i Krakauer afegeixen una sèrie més de premisses, algunes de les quals recollides d'altres autors (representa que estan fent una descripció de l'estat de la qüestió). Per exemple, un paral·lelisme entre la mecànica quàntica i la troballa estadística de patrons: P149:«a long-standing criticism of quantum mechanics is that it provides an effective means of calculation without providing conceptual understanding»[352]. L'estratègia argumentativa consisteix en utilitzar la mateixa crítica sobre un cos teòric amb gran acceptació i capacitat de predicció (com és la mecànica quàntica), sobre un altre cos teòric que precisament està en qüestió. L'argument es podria haver plantejat, de forma més clara, de la següent manera: així com no posem en dubte el coneixement i comprensió del món que aporta la mecànica quàntica (tot i que és antiintuïtiva en molts dels seus enunciats), no té sentit posar en dubte el coneixement i comprensió del món que aporten els LLM. El problema d'aquest argument, que proposicionalment és impecable i segurament acceptarien tots els *AI denials*, és que no defensa la tesi de l'article, ja que del que es predica comprensió és dels humans que interpreten el resultat matemàtic de la mecànica quàntica, no de la mecànica quàntica mateixa (ningú ha defensat fins ara que la mecànica quàntica tingui ment), mentre que l'argument inicial pretenia defensar que els LLM (com a conjunt de línies de codi) tenien una ment. Resulta curiós que Mitchell i Krakauer no senyalin aquesta fal·làcia.

Finalment, Mitchell i Krakauer enllacen un reguitzell de preguntes retòriques que els permetran tancar l'article amb una falsa aparença d'objectivitat. Pel seu interès dramàtic, s'exposa a continuació el paràgraf sencer:

> The key questions of the debate about understanding in LLMs are the following: (1) Is talking of understanding in such systems simply a category error, mistaking associations between language tokens for associations between tokens and physical, social, or mental experience? In short, is it the case that these models are not, and will never be, the kind of things that can understand? Or conversely, (2) do these systems (or will their near-term successors) actually, even in the absence of physical experience, create something like the rich concept-based mental models that are central to human understanding, and, if so, does scaling these models create ever better concepts? Or (3) If these systems do not create such concepts, can their unimaginably large systems of statistical correlations produce abilities that are functionally equivalent to human understanding? Or, indeed, that enable new forms of higher-order logic that

352 *Ibídem*, pàg. 5.

> humans are incapable of accessing? And at this point will it still make sense to call such correlations "spurious" or the resulting solutions "shortcuts?" And would it make sense to see the systems' behavior not as "competence without comprehension" but as a new, nonhuman form of understanding? These questions are no longer in the realm of abstract philosophical discussions, but touch on very real concerns about the capabilities, robustness, safety, and ethics of AI systems that increasingly play roles in humans' everyday lives.[353]

De fet, aquesta llista, que té certa forma de condicional niat, s'acabaria fàcilment responent afirmativament a la primera pregunta: sí, hi ha un error conceptual en el mer plantejament de predicar comprensió sobre una entitat mecànica o, més concretament, un conjunt de línies de codi, en la mesura que el concepte de comprensió s'aplica en organismes que estiguin vius. El sol fet d'enllaçar la resta de preguntes, una sota de l'altra, com obrint la porta a un bloc de possibilitats niades (i si no... llavors? i si no... llavors? i si no... llavors?), pretén traslladar cert grau de raonabilitat fins la més llunyana de les seves exclusions de possibilitat. Aquesta possibilitat remota permet, o creuen que permet, a Mitchell i Krakauer, plantejar una proposta etològica que pràcticament és l'inici d'una mitologia digital: P150:«It could thus be argued that in recent years the field of AI has created machines with new modes of understanding, most likely **new specie**s in a larger zoo of related concepts, that will continue to be enriched as we make progress in our pursuit of the elusive nature of intelligence»[354]. Per tant, més enllà de la multitud de sinònims que es podien haver escollit enlloc "zoo" per designar la llarga llista de conceptes relacionats, és evident que el plantejament que es fa, encara que sigui com a mera possibilitat argumental, és completament etològic.

Especialment per aquest últim article, però també perquè Mitchell és algú tècnicament irreprotxable, la seva contribució a la confecció d'una etologia digital és molt més perillosa, cosa que es pot il·lustrar a la taula de la següent forma:

353 *Ibídem*, pàgs. 5-6.

354 *Ibídem*, pàg. 6.

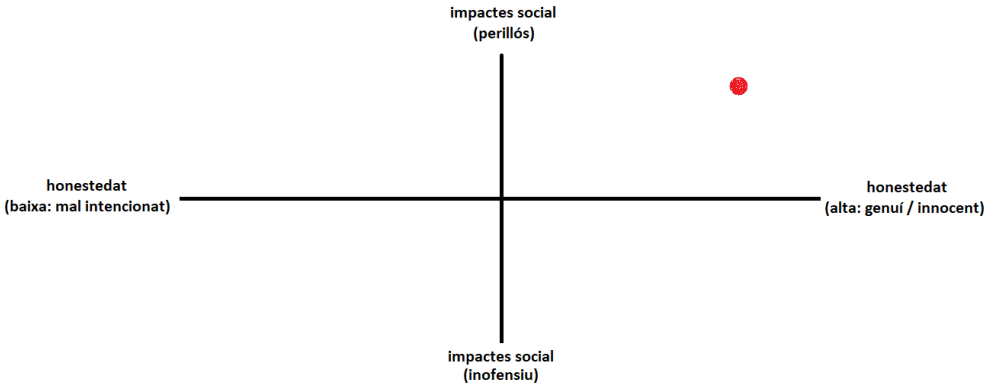

Figura 9: Nivell d'honestedat / impacte social de Melanie Mitchell

La intenció és fer notar una petita penalització en l'eix de l'honestedat tenint en compte el progressiu èmfasi en l'evolució i la possible mitificació final, però no es pot oblidar la contundència en la resposta a la carta de Musk i companyia. I també es fa notar que aquest tipus de discursos, per estar fonamentats més curosament, a la llarga poden tenir un impacte social més alt.

4.4 Rodney Brooks o com una altra digitalització és possible

Com ja s'ha anat veient a través de citacions puntuals en els apartats i també en els capítols anteriors, Rodney Allen Brooks (1954) pot exemplificar una mica la postura final d'aquest mostrari de propostes que col·laboren d'una manera o una altra en la confecció d'una etologia digital. Aquí es pretén sistematitzar breument l'estudi de la seva posició i molt concretament el pas entre el primer Brooks i el segon Brooks. Brooks és especialista en robòtica i actualment és professor emèrit del Massachusetts Institute of Technology. Va ser el fundador, president i CTO (*Chief Technology Officer*) de Rethink Robotics i actualment també és el CTO i cofundador de Robust AI, empresa que sembla que va fundar amb Gary Marcus (tot i que actualment no s'ha trobat cap referència a Marcus en el web de l'empresa). Va ser director del MIT Computer Science & Artificial Intelligence Laboratory (CSAIL). Es va llicenciar en Matemàtiques a la Flinders University of South Australia i es va doctorar en Informàtica a la Universitat de Stanford[355]. En poques paraules, és un expert en digitalització tant pels seus estudis com per la seva llarga experiència acadèmica (docent, investigadora i directiva) i empresarial (no sempre amb èxit, com ell mateix assumia[356]).

355 "Rodney Brooks – Roboticist" en *MIT Computer Science and Artificial Intelligence Laboratory*. Consultat el 16 de juliol de 2024 a: https://people.csail.mit.edu/brooks/

356 ZORPETTE, Glenn (entrevistadora); BROOKS, Rodney (entrevistat) (17.05.2023). "Just Calm Down About GPT-4 Already" en *IEEE Spectrum*. Consultat el 16 de juliol de 2024 a: https://spectrum.ieee.org/gpt-4-calm-down

S'identifica amb el primer Brooks allò que aquest defensa en articles com "Elephants Don't Play Chess" (1990) o "The relationship between matter and life" (2001), mentre que es considera que el segon Brooks queda ben definit pels comentaris que fa en una xerrada del 13 de maig de 2019 conduïda per ell mateix amb alguns col·legues seus com Seth Lloyd, Frank Wizek o el mateix Chalmers; per una entrada al seu blog el 23 de març de 2023 titulada "What Will Transformers Transform"; i per una entrevista que li fa Glenn Zorpette el 17 de maig de 2023 titulada "Just Calm Down About GPT-4 Already" a *IEEE Spectrum*.

El primer Brooks

El primer Brooks defensa que per construir una IA cal seguir els passos de l'evolució, com també ho fa Dreyfus, Marcus o Mitchell, però se li critica que volgués anar massa ràpid. En "Elephants Don't Play Chess" Brooks fa una presentació del que es pot arribar a fer en robòtica amb les noves tècniques (*nouvelle AI*) si enlloc d'utilitzar una programació simbòlica (entesa com una aproximació *top-down*), es construeixen robots amb sensors que capten l'entorn i que després s'articulen unificadament (*bottom-up*). El text –que és una defensa d'aquesta nova aproximació a la qual explícitament vol diferenciar d'unes altres propostes innovadores conegudes com xarxes neuronals (la base del la IA generativa actual)– sembla un mostrari del que han fet en el laboratori del MIT durant la dècada dels vuitanta, talment com si l'objectiu fos aconseguir finançament (en aquells moments, entre 1987 i 1993, es vivia un dels hiverns cíclics en els quals entra la investigació en IA[357]).

La tesi principal de la nova aproximació és un postulat emergentista: P151:«The new methodology bases its decomposition of intelligence into individual behavior generating modules, whose coexistence and co-operation let more complex behaviors emerge»[358], i com a postulat que és, és un acte de confiança: P152:«Nouvelle AI relies on the emergence of more global behavior from the interaction of smaller behavioral units»[359]. Un dels defensors d'un emergentisme fort és Chalmers, que defensa que la consciència n'és un exemple. Per tant, tot i que aquí queda clar que Brooks no està parlant de cap tipus de fenomen mental, sí que hi ha un ús intencionat del terme que il·lumina o distorsiona l'explicació (era fàcil evitar el verb i simplement dir que s'espera que el

357 FRANCESCONI, Enrico (2022). "The winter, the summer and the summer dream of artificial intelligence in law: Presidential address to the 18th International Conference on Artificial Intelligence and Law" en *Artificial Intelligence and Law*, 30 (3). Consultat el 16 de juliol de 2024 a: https://doi.org/10.1007/s10506-022-09309-8

358 BROOKS, Rodney A. (1990). "Elephants Don't Play Chess" en *Robotics and Autonomous Systems*, 6 (1990), pàg. 3. Consultat el 16 de juliol de 2024 a: https://people.csail.mit.edu/brooks/papers/elephants.pdf

359 *Ibídem*, pàg. 12.

robot apliqui les funcionalitats programades per resoldre situacions no previstes). Per tant, es pot considerar això un petit plantejament etològic digital.

Més endavant, tanmateix, l'evidència d'estar perseguint un etologia digital es fa clara si s'encadenen les següents cinc proposicions:

> P153: «Given that neither classical nor nouvelle AI seem close to revealing the secrets of the holy grail of AI, namely general purpose human level intelligence equivalence [...]».[360]
>
> P154:«Charmingly, it has been hoped that intelligence will somehow emerge from these simple numeric computations carried out in the sea of symbols».[361]
>
> P155:«We already have an existence proof of the possibility of intelligent entities — human beings. Additionally many animals are intelligent to some degree (...). They have evolved over the 4.6 billion year history of the earth […]». [362]
>
> P156:«It is instructive to reflect on the way in which earth-based biological evolution spent its time».[363]
>
> P157:«That essence is the ability to move around in a dynamic environment, sensing the surroundings to a degree sufficient to achieve the necessary maintenance of life and reproduction. This part of intelligence is where evolution has concentrated its time—it is much harder. This is the physically grounded part of animal systems».[364]

Per P153 se sap que l'objectiu final és fer robots que actuïn amb intel·ligència equivalent a la humana i per fer-ho Brooks proposa explícitament copiar l'evolució (P155-P157). Si els sistemes vius s'han passat 3,5 bilions d'anys per aprendre a reaccionar a l'entorn –i aquí Brooks incorpora un interessant còmput que posa sobre la taula que els hominins van aparèixer fa només 2,5 milions d'anys i que l'agricultura es va inventar (potser "començar a practicar" seria més curós) fa 19 mil anys, l'escriptura en fa 5 mil i el coneixement expert (*"expert" knowledge*) fa poques centenes– per què no es dediquen els esforços en programar robots que bàsicament sàpiguen reaccionar, es pregunta. Potser així, també apareixerà màgicament la intel·ligència (P154).

És cert que alguns termes que utilitza Brooks per descriure els dos plantejaments (tant el simbòlic com el *nouvelle*) tenen un marcada patina religiosa (no només utilitza *rely* per descriure aquesta confiança cega en un o altre projecte –P158:«The user again is relying on expectations without hard proofs»[365]–, sinó també *faith*), però també és cert que en algun moment sembla un plantejament basant en un "i tu més" orientat a aconseguir un permís o un augment de pressupost:

360 *Ibídem*, pàg. 4.

361 *Ibídem*, pàg. 5.

362 *Ídem*.

363 *Ídem*.

364 *Ídem*.

365 *Ibídem*, pàg. 12.

P159:«But just as the symbol system people are allowed to work incrementally in their goals, so should the physical grounding people be allowed»[366]. Ara bé, això no condiciona que apliqui un argumentari evolucionista (comportament 10 d'una etologia digital raonada), ni tampoc que resulti curiós (per no dir lleugerament superb) pretendre emular un procés com l'evolució en uns pocs anys (encara que siguin centenes, que temporalment quadraria amb aquest període de "coneixement expert" que singularitza, talment com si abans no n'hi hagués hagut de coneixement expert).

Onze anys més tard, Brooks segueix pràcticament igual: la robòtica s'ha d'inspirar en la vida i com en aquesta la intel·ligència ha sorgit de la matèria. Tanmateix, en "The relationship between matter and life" (2001), el verb "emergir" només apareix una vegada, en sentit figurat i no per referir-se a res vinculat amb el projecte de la IA. Aquí Brooks parla explícitament d'inspiració: P160:«The disciplines of artificial intelligence and artificial life build computational systems inspired by various aspects of life»[367]. Aquesta inspiració es converteix en una comprensió, i aquesta comprensió té un objectiu pràctic, tant per una disciplina com per l'altra: P161:«Researchers in artificial intelligence (AI) and artificial life (Alife) are interested in understanding the properties of living organisms so that they can build artificial systems that exhibit these properties for useful purposes»[368]. Per tant, tot i que els noms propis són els mateixos (i.e. evolució, naturalesa, organismes vius, entre d'altres), la forma de tractar-los és diferent: *emergir* s'ha convertit en *inspirar* i *inspirar* en *entendre* per *utilitzar*. Aquest és el primer pas cap el segon Brooks.

De fet, l'objectiu de l'article és aportar noves perspectives davant l'estancament evident del sector i la conseqüent decepció: P162:«At the heart of this disappointment lies the fact that neither AI nor Alife has produced artefacts that could be confused with a living organism for more than an instant»[369]. I això es deu a què quelcom està fallant: P163:«But we are not good at modelling living systems, at small or large scales. Something is wrong»[370]. Però, mentre el primer Brooks encara batega i té una resposta –cal un gir que deixi tanta teoria poc fonamentada i es centri en aspectes més enginyerils, sense renunciar a l'objectiu de modelar sistemes vius, expressió que podria confondre's perfectament amb alguna de Mary Shelley com aquesta: «Pursuing these reflections, I thought that if I could bestow animation upon lifeless matter, I might in process of time (although I

366 *Ídem.*

367 BROOKS, Rodney (18.01.2001). "The relationship between matter and life" en *Nature*, vol. 409, pàg. 409. Consultat el 17 de juliol de 2024 a: https://www.nature.com/articles/35053196

368 *Ídem.*

369 *Ídem.*

370 *Ibídem*, pàg. 410.

now found it impossible) renew life where death had apparently devoted the body to corruption»[371]– el segon Brooks, que comença a emergir, enlloc de proposar un nova solució, comença dubtar si el que cal es refundar el problema. Tanmateix, el 2001 encara té certa esperança (i certa ingenuïtat) com es nota al plantejar-se només com a possibilitat el fet que potser encara no se'n sap prou del tema: P164:«Building models that are below some complexity threshold also would mean that there is nothing in principle that we do not understand about intelligent or living systems»[372]. Tot i que una idea (que acabarà expressant més clarament el segon Brooks) ja li ronda pel cap: P165:«We would then need to find new ways of thinking about living systems to make any progress, and this will be much more disruptive to all biology»[373]. Ara bé, el 2001 Brooks encara no està allà, i es rebelluga entre propostes cap de les quals l'acaba de convèncer: la proposta de Roger Penrose, de que la consciència sigui un efecte quàntic en les cèl·lules nervioses, no està encara ben teoritzada; la idea de Chalmers, que ha d'haver-hi una propietat física de les partícules, l'equipara amb l'alè vital o l'ànima; la invenció de noves matemàtiques generatives més que descriptives, sembla desencaminada (*misguided*)... Brooks és conscient que hi ha alguna cosa que no quadra i comença a plantejar una hipòtesi: P166:«An analogy to the sort of thing that might be missing is computation — not as the undiscovered feature itself but as an analogy for the type of thing we might be looking for»[374]. D'aquesta idea, de que la metàfora computacional no dona més de si, naixerà el que en aquest treball s'anomena el segon Brooks.

El segon Brooks

Una mostra de que els dubtes del 2001 s'han convertit en certesa és la conversa o, literalment, taula rodona que Brooks grava i publica amb alguns dels seus companys la llista dels quals és la següent: Seth Lloyd, Frank Wilczek, Neil Gershenfeld, Stephen Wolfram, W. Daniel Hillis, John Brockman, Tom Griffiths, Peter Galison, David Chalmers, Carlonie Jones, Geroge Dyson, Alison Gopnik i Freeman Dyson (segons ordre d'intervenció). L'acte apareix transcrit en un article publicat a *Edge* i titulat "The Cul-de-Sac of the Computational Metaphor" el 13 de maig de 2019.

Brooks fa una breu introducció històrica per acabar plantejant el problema que vol tractar amb els seus companys: P167:«Neuroscience uses computation as a metaphor, and I question whether

371 SHELLEY, Mary (1818). *Frankenstein; or, The Modern Prometheus*, Cambridge (Massachusetts), The MIT Press, 2017, pàgs. 36-37. Consultat el 17 d'agost de 2024 a: https://rauterberg.employee.id.tue.nl/lecturenotes/DDM110%20CAS/Shelley-1818%20Frankenstein.pdf

372 BROOKS, Rodney (18.01.2001). "The relationship between matter and life" en *Nature*, vol. 409, pàg. 410. Consultat el 17 de juliol de 2024 a: https://www.nature.com/articles/35053196

373 *Ídem.*

374 *Ibídem*, pàg. 411.

that's the right set of metaphors»[375]. I insisteix quan li puntualitzen que el processament de la informació quàntica (*quantum information processing*) també funciona, tot i que ineficientment, en computació clàssica: P168:«My point is that I don't think that classical computation is the right mechanism to think about quantum mechanics. There are other metaphors»[376]. I, per tant, el dubte de 2001 s'ha convertit en una certesa el 2019: cal refundar el camp i una de les obres a tenir en compte és *Metaphors We live By*, de Lakoff i Johnson's. I insisteix en que la computació no pot ser l'única manera de traduir el procés de pensar: P169:«Is information processing the right metaphor there? Or are control theory and resonance and synchronization the right metaphor? We need different metaphors at different times, rather than just computation»[377]. Perquè la computació no deixa de ser una simplificació programable: P170:«We have fairly simple dynamics in our computational spaces because that's what we can generate with computation»[378]. I explica el motiu pel qual, al seu parer, s'ha restat tant de temps en una metàfora que, des de diverses perspectives, ja es veia que no tenia més recorregut, que era una simplificació: P171:«The reason for why we got stuck in this cul-de-sac for so long was because Moore's law just kept feeding us, and we kept thinking, "Oh, we're making progress, we're making progress, we're making progress." But maybe we haven't been»[379]. I Brooks acaba reivindicant que ell, el 2001, en un article a *Nature* ja ho havia dit, però que ningú li havia fet cas.

Més endavant [23:00], Gershenfeld li demana que digui quelcom en positiu, que proposi alguna cosa ell que hi ha pensat tant. Brooks comença aquí a plantejar dues idees, una de negativa, que és la que acabarà configurant el segon Brooks, i una de positiva, que sembla encara un espeternec del primer Brooks, tot i que és cert que la porten a col·lació alguns dels companys de taula. Començant per la segona, Brooks observa que la computació dona bons resultats en la predicció, però que els sistemes biològics no funcionen per predicció sinó per adaptació; per tant, cal una computació adaptativa. Chalmers aquí [36:50], interpel·lat directament, intervé per dir que el *machine learning* és computació adaptativa, cosa que no sembla acabar de convèncer a Brooks, que li resumeix com funciona la computació: P172:«Let me give you an example that fits your model there. We went from the Turing machine to the RAM model, and current computational

375 BROOKS, Rodney A. (13.05.2019). "The Cul-de-Sac of the Computational Metaphor" en *en Edge*, pag. 2. Consultat el 17 de juliol de 2023 a: https://www.edge.org/conversation/rodney_a_brooks-the-cul-de-sac-of-the-computational-metaphor

376 *Ídem.*

377 *Ibídem*, pàg. 3.

378 *Ibídem*, pàg. 4.

379 *Ibídem*, pàg. 5.

complexity is really built on the RAM model of computation. It's how space and time trade off in computation»[380]. Per tant, com sintetitza Wolfram, el problema de base és la diferència entre continu (analògic) versus discret (digital). I Brooks insisteix, insistència que dibuixa molt bé aquest impàs entre el primer Brooks i el segon: P173:«There may be something somewhat different from that that we just haven't seen yet in the large system of lots of processes happening without clear interfaces, and lots of statistical stuff going on—statistical just because you don't know everything»[381]. Aquest "hi ha d'haver d'alguna forma alguna cosa diferent..." és un primer pas per arribar a aquest segon Brooks, a qui ajuda a expressar-se el seu company Lloyd:

> BROOKS: Yes. All of us here would be terribly surprised if we're at the beach and we saw a robot dolphin come out of the water that had been built by dolphins. We just don't expect dolphins to have the cognitive capability to do what we're trying to do in artificial intelligence. We don't think they have it, nor the dexterity.
>
> LLOYD: We expect them to have better sense than to do such a thing.
>
> BROOKS: Yes [...][382]

Aquí es defensa que aquest "Sí", amb tots els dubtes encara revoltant pel cap, és la primera expressió plena del segon Brooks.

Una segona expressió d'aquest viratge es troba en una entrada al seu blog el 23 de març de 2023 titulada "What Will Transformers Transform?" (*Generative Pre-trained Transformer models* o GPTs). Una de les primeres idees del text és que, tot i la sobrevaloració d'aquesta tecnologia, poden ser una eina útil: P174:«In short, there will be valuable tools produced, and at the same time lots of damaging misuse»[383]. Per tant, el tractament és d'eina i, com qualsevol eina, es pot fer servir bé o malament, que no vol dir que en si mateixa sigui neutre, doncs si l'ús majoritari acaba sent per fer un mal, llavors no és un tema d'ús, sinó d'ús condicionat. En qualsevol cas, Brooks només entrarà ara en aquestes consideracions al fer un llistat de 9 prediccions del que passarà entre el moment actual i el 2030 amb l'ús del GPT: recuperació del valor de la Wikipedia com a coneixement fet per humans; la mercantilització d'apps molt específiques usant GPT; persistència de les al·lucinacions; no serà aprofitable seriosament en robòtica; serà més fàcil crear nous programes iguals que els

380 *Ibídem*, pàg. 8.

381 *Ibídem*, pàg. 9.

382 *Ídem.*

383 BROOKS, Rodney (23.03.2023). "What Will Transformers Transform?" en *Robots, AI, and other stuff.* Consultat el 17 de juliol de 2024 a: https://rodneybrooks.com/what-will-transformers-transform/

actuals; hi haurà un creixement d'empreses de certificació de drets d'autor; es construiran coses amb les quals ningú ha pensat encara; creixerà la desinformació; i hi haurà una nova categoria pornogràfica. Per tota la resta, Brooks té clar que els GPTs són eines a les quals no es poden aplicar atributs humans i el motiu: P175:«GPT-n cannot reason, and it has no model of the world. It just looks at correlations between how words appear in vast quantities of text from the web, without know how they connect to the world. It doesn't even know there is a world»[384]. I que el que dona sentit o tot plegat és la mirada humana que relliga les respostes: P176:«Many successful applications of AI have a person somewhere in the loop. Sometimes it is a person behind the scenes that the people using the system do not see, but often it is the user of the system, who provides the glue between the AI system and the real world»[385] – aquesta afirmació recorda molt la de Kate Crawford quan afirma: «In contrast, in this book I argue that AI is neither artificial nor intelligent»[386]. Brooks afegeix també una reflexió sobre la tecnologia que se li atribueix a Roy Amara i que se sol anomenar la Llei d'Amara: P177:«We tend to overestimate the effect of a technology in the short run and underestimate the effect in the long run»[387]. Aquesta idea allunya a Brooks encara més d'una possible etologia digital: la tecnologia és tecnologia i té un efecte estudiable sobre la població, com va fer Lewis Mumford a *Technics & Civilization* (1934)[388] o, més específicament, a Jacques Ellul a *La technique ou l'enjeu du siècle* (1954)[389], qui defensava que cada nova tecnologia crea uns problemes que requeriran més tecnologia per resoldre'ls, en un espiral sense fi que oprimeix la humanitat.

Per tancar l'anàlisi etològic d'aquest segon Brooks (o de manca d'etologia), s'analitzarà un últim text, l'entrevista que li va fer Glenn Zorpette per *IEEE Spectrum* el 17 de maig de 2023. Aquí Brooks defensa que les eines han de ser útils per resoldre problemes humans, però que això no vol dir que hagin de ser igual que els humans: simplement cal que tinguin en compte la seva existència i necessitats. També fa una explicació sobre quan funciona una innovació tecnològica: requereix d'un canvi en l'entorn.

---

384 *Ídem.*

385 *Ídem.*

386 CRAWFORD, Kate (2021). *Atlas of AI. Power, Politic, and the Planetary Costs of Artificial Intelligence*, New Haven, Yale University Press, 2021, pàg. 8.

387 BROOKS, Rodney (23.03.2023). "What Will Transformers Transform?" en *Robots, AI, and other stuff.* Consultat el 17 de juliol de 2024 a: https://rodneybrooks.com/what-will-transformers-transform/

388 MUMFORD, Lewis (1934). *Técnica y civilización*, Logroño, Pepitas de calabaza S.L., 2020.

389 ELLUL, Jacques (1954). *La technique ou l'enjeu du siècle*, Paris, 1954 (trad. cast.: *La edad de la técnica*, Barcelona, Ediciones Octaedro, 2003).

Brooks comença fent una diferència entre actuació i competència que va al moll de l'os:

> P178: If I can just expand on that a little. When we see a person with some level performance at some intellectual thing, like describing what's in a picture, for instance, from that performance, we can generalize about their competence in the area they're talking about. And we're really good at that. Evolutionarily, it's something that we ought to be able to do. We see a person do something, and we know what else they can do, and we can make a judgement quickly. But our models for generalizing from a performance to a competence don't apply to AI systems.[390]

Aquí, l'evolució no serveix ni per inspirar-s'hi ni per emmirallar-s'hi, sinó per explicar com els humans traslladem per inèrcia un capacitat antropològica a un sistema d'IA. L'explicació recorda la que Marcus donava també en el *Rebooting* del mateix fenomen i que allà anomena la bretxa de la credulitat (*the gullibility gap*): «We attribute intelligence to computers because we have evolved and lived among human beings who themselves base their actions on abstractions like ideas, beliefs, and desires»[391]. Per tant, segueix Brooks, no és que el GPT respongui amb sentit, sinó que el text que retorna sona com si tingués sentit: P179:«What the large language models are good at is saying what an answer should sound like, which is different from what an answer should be»[392]. Això no vol dir que els LLM no siguin útils: poden facilitar la recerca d'informació a un usuari final que desconegui el llenguatge informàtic.

La part més interessant de l'entrevista és quan Brooks explica per què no funcionaran els cotxes automàtics si no es fan més canvis: mai abans ha funcionat un nou mitjà de transport sense adaptar la infraestructura per la qual es desplaça:

> P180: And as a species, humanity, we have changed up our mobility infrastructure multiple times. In the early 1800s, it was steam trains. We had to do enormous changes to our infrastructure. We had to put flat rails right across countries. When we started adopting automobiles around the turn from the 19th to the 20th century, we changed the roads. We changed the laws. People could no longer walk in the middle of the road like they used to. We changed the infrastructure. When you go from trains that are driven by a person to selfdriving trains, such as we see in airports and a few out there, there's a whole change in infrastructure so

390 ZORPETTE, Glenn (17.05.2023). "Just Calm Down About GPT-4 Already And stop confusing performance with competence, says Rodney Brooks" en *IEEE Spectrum*. Consultat el 18 de juliol de 2024 a: https://spectrum.ieee.org/gpt-4-calm-down

391 MARCUS, Gary; DAVIS, Ernest (2019). *Rebooting AI. Building Artificial Intelligence We Can Trust*, Nova York, Vintage Books (Penguion Random House LLC), 2020, pàg. 18.

392 *ZORPETTE, Glenn (17.05.2023). "Just Calm Down About GPT-4 Already And stop confusing performance with competence, says Rodney Brooks" en IEEE Spectrum*. Consultat el 18 de juliol de 2024 a: https://spectrum.ieee.org/gpt-4-calm-down

> that you can't possibly have a person walking on the tracks. We've tried to make this transition [to self-driving cars] without changing infrastructure. You always need to change infrastructure if you're going to do a major change.[393]

La reflexió denota un coneixement de la història de la tecnologia que, novament, recorda a Mumford, un dels plantejaments més allunyats d'una etologia digital, i a Weizenbaum, quan afirma que hi ha altres maneres de fer tecnologia:

> El ordenador del avión puede presentarse como ejemplo de Inteligencia Artificial. Pero no tiene nada que ver con la simulación del pensamiento humano. Resulta que las palas mecánicas pueden hacer una zanja de forma más rápida que los humanos. Los ordenadores pueden hacer cosas mucho más rápidas que nosotros, sobre todo cosas matemáticas, incluso pueden ser más sensibles que los humanos. El más leve cambio de dirección que tome el aeroplano, tan leve que no lo nota un ser humano, se percibe mediante los giroscopios. Además, uno de estos sistemas puede aterrizar un avión de forma más suave que un piloto.[394]

Aquest és l'enfocament que acaba plantejant Brooks, un enfocament merament tècnic i que queda plasmat en el concepte "robot col·laboratiu". Aquestes màquines estaran dissenyades tenint en compte que han d'interactuar amb humans: P181:«So we're trying to make our robots human-centered, we call it. They're aware of people. They're using convolutional neural networks to see that that's a person, to see which way they're facing, to see where their legs are, where their arms are»[395]. Només l'expressió "són conscients de la gent" podria ser considerada etològica, però en el context de maquinària per magatzem, queda completament descartada. Tan descartada quan, a més a més, no tenen forma humanoide:

> P182: But then the magic of our robot is that it looks like a shopping cart. It's got handlebars on it. If a person goes up and grabs it, it's now a powered shopping cart or powered cart that they can move around. So [the warehouse workers] are not subject to the whims of the automation. They get to take over. When the robot's clearly doing something dumb, they can just grab it and move it, and it repairs.[396]

393 *Ídem.*

394 WEIZENBAUM, Joseph (1992). "Entrevista a Joseph Weizenbaum" en *Telos*, núm.38, Fundación Telefónica. Consultat el 16 d'agost de 2024 a: https://telos.fundaciontelefonica.com/archivo/numero038/entrevista-a-joseph-weizenbaum/

395 *Ídem.*

396 *Ídem.*

Quan un robot té forma de carretó d'anar a comprar, es pot espatllar i cal reparar-lo, i l'únic antropocèntric que cal tenir en compte és que no atropelli humans, es pot afirmar que aquest plantejament no té cap element etològic.

Per tots aquests motius, si s'hagués de representar gràficament el nivell etològic de Brooks, especialment d'aquest segon Brooks, es podria il·lustrar així:

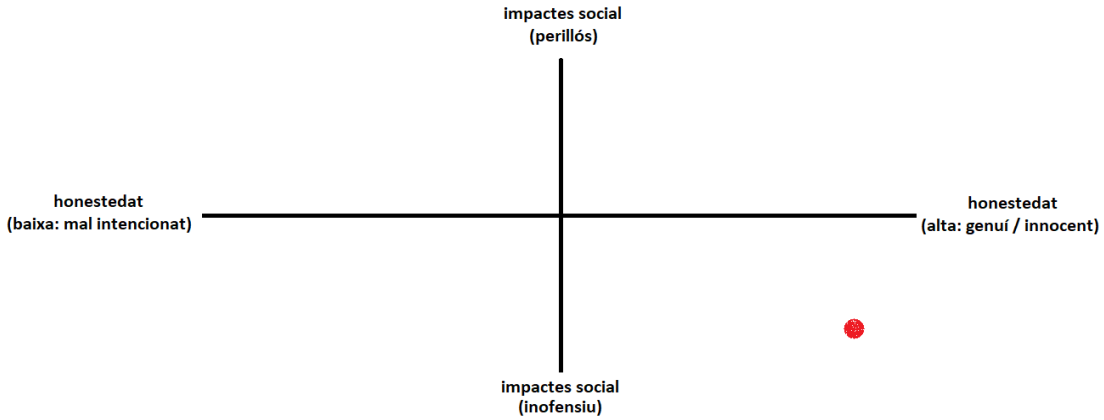

Figura 10: Nivell d'honestedat / impacte social de Rodney Brooks

N.B. Brooks va publicar una nova entrada al seu blog el 8 de desembre de 2023. No caldria justificar per què no s'ha utilitzat aquesta entrada més recent, ja que que aquest treball no pretén ser una relació exhaustiva, sinó un mostrari ordenat d'etologies digitals (per tant, la tria de textos està al servei del relat). Ara bé, la nova entrada es titula "Three Things that LLMs Have Made Us Rethink" (i per *Us* vol dir *Me*), i, en la mesura que podria semblar que s'intenta amagar un canvi en Brooks que afacteria al mateix relat, cosa que no passa, sembla necessari comentar-lo.

1. Brooks no fa cap afirmació que sigui etològica, però sí que reconeix que mai hagués pensat que solament amb una sintaxis fos possible aconseguir un sistema d'intercanvi de frases tan ben aconseguit (*syntax suffices* és com es coneix la postura que defensava William Rapaport el 1988 contra Searle: «So, my thesis is that (suitable) purely syntactic symbol-manipulation of the system's knowledge base (its "mind") suffices for it to understand natural language»[397]).

2. Brooks descarta que el Test de Turing tingui cap sentit quan les empreses es preparen específicament per superar-lo (Dennett ja ho havia comentat el 1988 amb l'exemple de la "gran ciutat"), i per això no li dona cap valor que el ChatGPT sembli que ho faci. Ara bé, reconeix que

397 RAPAPORT, Willliam J. (1988). "Syntactic Semantics: Foundations of Computational Natural-Language Understanding" en *Aspects of Artificial Intelligence*, James Fetzer (ed.), Kluwer Academic Publishers, Dordrecht, 1988, pàgs. 85-86.

l'exemple de l'habitació xinesa de Searle ha quedat superat – Searle havia defensat que el significat el donava els ulls del que mirava, en el mateix sentit que Brooks apel·lava a la cola que relliga tot sistema d'IA: «The information in the Chinese case is solely in the eyes of the programmers and the interpreters, and there is nothing to prevent them from treating the input and output of my digestive organs as information if they so desire»[398]. Brooks accepta que ara es fa evident que un sistema de traducció de símbols pot construir frases fins a tal punt que sembli que les entengui (i aquest és l'únic element que tindria algun matís etològic).

3. Brooks creu que la teoria de Chomsky sobre la gramàtica universal queda seriosament qüestionada: «On the other hand the ability to learn human grammar with no mechanism for grammar built in is certainly a surprise, at least to time traveling AI researchers from thirty or even twenty years ago»[399]. Aquesta idea és la que Shanahan posarà sobre la taula en l'article que servirà de pont entre la primera part d'aquest treball i la segona.

4. Brooks acaba demanant més temps (torna a estar desconcertat): «My old self and my today self are not being intellectually coherent, so I am going to have to think about this some more over the next few years and refine, perhaps rethink, but certainly change in some way what it is I conclude from both Searle and ChatGPT existing».[400]

398 SEARLE, John R. (1980). "Mind, brains, and programs" en *Behavioral and Brain Sciences*, 3 (3), Cambridge University Press, pàg. 420.

399 BROOKS, Rodney (08.12.2023). "Three Things That LLMs Have Made Us Rethink" en *Robots, AI, and other stuff*. Consultat el 17 de juliol de 2024 a: https://rodneybrooks.com/three-things-that-llms-have-made-us-rethink/

400 *Ídem.*

Intermezzo

Conclusions sobre els autors escèptics i els autors de la por

En el camp de la intel·ligència artificial no hi ha autors realment escèptics: si hom dedica la seva vida a aquesta investigació és perquè confia en la possibilitat que, tard o d'hora, es pugui crear realment una intel·ligència artificial. L'única diferència entre els autors que aquí s'han denominat "de la por" i els "escèptics" és metodològic i de grau: mentre que els autors de la por utilitzen estratègies centrades en la imminència de la IA, els escèptics la situen a llarg termini; mentre que els autors de la por situen aquesta IA per sobre de la intel·ligència humana, els escèptics o bé són conscients que està molt per sota (al nivell de programar formigues robòtiques) o, simplement, en una altra branca evolutiva diferent de la dels hominins. Per tant, d'una forma més o menys elaborada, tots acabaran col·laborant en la confecció d'una etologia digital, en la mesura que el certificat d'haver aconseguit una intel·ligència artificial només l'obtindran en el moment en què puguin situar aquesta entitat entremig dels organismes amb els quals es comparen. En aquell moment, aquesta entitat mecànica, aquest conjunt d'algoritmes, esdevindrà, des del seu punt de vista, un organisme, com a mínim, amb una ment i, en cert sentit, amb vida: el somni del Dr. Frankenstein actualitzat.

George Zarkadakis, expert en innovació en Intel·ligència Artificial i Dades[401], defensa que aquest somni té un origen genètic: la capacitat dels humans moderns per suposar una ment en els altres, fins i tot, quan aquests altres no són humans (com els animals) o són objectes (per exemple, els tòtems). D'aquesta capacitat de projectar en els altres la pròpia consciència, sortiria també el desig d'imaginar com s'ha adquirit la pròpia consciència, responent-se aquesta pregunta de forma diferent a partir de les metàfores de cada època: «First came mud, then water or humours, then mechanics, the electric current or spark of life, followed by the telegraph and now the computer. For each of these metaphors, people have imagined automata, artificial artefacts set in motion by technologies that support the metaphors»[402]. Per tant, per entendre què hi ha a la base d'una etologia digital cal estudiar la seva base metafòrica, en aquest cas, en aquesta època, de la metàfora computacional.

401 George Zarkadakis. En *Linkedin*. Consultat el 26 d'agost de 2023 a: https://www.linkedin.com/in/gzarkadakis

402 ZARKADAKIS, George (2015). *In our own image*, Londres, Rider Books, 2015, pàg. 44.

Perdre's en les dreceres

L'ús de metàfores per parlar dels LLM s'ha tornat habitual. Depenent de com es faci, aquest ús pot contribuir a una etologia digital. No sempre aquestes metàfores són explícites: com s'ha vist abans, a vegades són meres intencions que el programador s'imposa abans de començar un bucle que anomena UNDERSTAND, com ironitzava McDermott ja el 1976. A vegades, són dreceres (*shortcuts*) que fan els LLM per etiquetar una imatge, com denuncia Mitchell el 2021. Altres són abreviacions (*shorthands*) que es fan servir i que, fora de context, poden ser malinterpretades. Això és el que defensa Murray Shanahan en l'article titulat "Talking About Large Language Models"[403]: hi ha abreviacions que no surten a compte, i una d'elles és la d'extrapolar el llenguatge que es fa servir en un entorn tècnic entre especialistes de la matèria, a un entorn divulgatiu. Com s'ha descrit abans, aquest canvi d'ubicació d'un discurs és una de les formes per generar en el públic la idea que la IA és com una nova espècie, amb un comportament autònom: es prenen termes amb un significat molt concret en un camp i s'extrapolen en un altre més general, cosa que fa canviar parcialment el seu significat.

Shanahan creu que això és el que està passant arran de la proliferació d'aplicacions que utilitzen models de llenguatge gran (*Large Language Models* en anglès, d'aquí l'acrònim LLM), com GPT, BERT i LaMDA, o BlenderBot (propietat d'OpenAI, Google i Meta respectivament). Els LLM són models matemàtics de distribució estadística que, a partir d'un conjunt de textos, troben una sèrie de patrons entre tots els caràcters (*token*) que componen aquests textos. Aquests patrons els permeten predir amb molt d'encert quines són, donada una frase, les paraules més probables que vindran a continuació. Tot i saber que aquest encert es deu a què el conjunt de textos amb els quals s'han desenvolupat aquests LLM és elevadíssim, quan es fan servir, ben bé fa la impressió que saben què es diuen.

Entre les paraules que Shanahan considera que s'estan malinterpretant hi ha els verbs que tenen càrrega psicològica, com "sap" (*knows*), "creu" (*believes*) o fins i tot "pensa" (*thinks*), aquest darrer verb d'ús molt comú per descriure quan un procés informàtic triga una estona en retornar el resultat. Shanahan sosté, mesos abans de la famosa carta pública "Pause Giant AI Experiments: An Open Letter"[404], que en el context dels LLM l'ús d'aquests verbs és perillós, ja que porta un públic

403 SHANAHAN, Murray (16/02/2023). "Talking About Large Language Models" en arXiv preprint arXiv:2212.03551. Consultat el 8 d'abril de 2023 a: https://arxiv.org/pdf/2212.03551.pdf

404 "Pause Giant AI Experiments: An Open Letter" (22/03/2023). Consultat el 8 d'abril de 2023 a: https://futureoflife.org/open-letter/pause-giant-ai-experiments/. L'article de Shanahan va ser enviat a revisió el desembre de 2022, mentre que la carta es va fer pública el 22 de març de 2023, curiosament, a hores d'ara, encara sense la seva signatura.

no especialista a pensar que la IA fa allò que realment no fa -és a dir, pensar, creure o saber-, i que cal posar fre a aquesta confusió abans de seguir alliberant nous productes al mercat: «To mitigate this trend, this paper advocates the practice of repeatedly stepping back to remind ourselves of how LLMs, and the systems of which they form a part, actually work».[405]

Per explicar exactament què fa un LLM, Shanahan tradueix la pregunta que se sol fer a qualsevol d'aquests xats automàtics de la següent manera: enlloc de pensar que quan s'està preguntant per "qui va ser la primera persona que va caminar sobre la Lluna", el sistema està buscant la resposta en alguna enciclopèdia *online* (cosa que implicaria, per altra banda, que el LLM ha entès la pregunta) tal i com faria qualsevol humà en aquesta situació si no conegués la resposta, cal entendre que el sistema el que fa és buscar com acabar la frase que comença per "la primera persona que va caminar sobre la Lluna va ser...", tal i com fa el predictor de text del mòbil quan escrivim "bona" i proposa "nit". En concret, Shanahan explicita de la següent forma com s'executa la pregunta inicial en el LLM: «Given the statistical distribution of words in the vast public corpus of (English) text, what words are most likely to follow the sequence "The first person to walk on the Moon was "? A good reply to this question is "Neil Armstrong"»[406]. Per tant, el procediment que realitza un LLM no té res a veure amb el procediment humà de resolució, tot i que el resultat és el mateix: no hi ha cap relació entre la resposta i la realitat externa al text. D'aquí part de la confusió: s'apliquen verbs que tenen un significat suficientment clar en un context humà per descriure procediments tecnològics insuficientment clars o, com a mínim, que no encaixen bé amb el significat habitual d'aquest verbs; en altres paraules, en contextos tecnològics cal entendre que aquests verbs tenen un ús metafòric.

L'ús d'un llenguatge metafòric, ja sigui per escurçar explicacions ja sigui per explicar conceptes difícils de comprendre, ha estat una pràctica habitual i l'exemple més il·lustratiu és segurament alguns dels mites, símils o al·legories que fa servir Plató en els diàlegs. Shanahan també admet que és natural fer servir un vocabulari antropomòrfic per parlar d'artefactes i que això no genera cap problema si tots els parlants són conscients que es tracta d'abreviacions útils per no haver d'estar fent especificacions llarguíssimes: són dreceres. Cada dia tots podem utilitzar expressions com les següents sense caure en cap equívoc: «My watch doesn't realise we're on daylight saving time. My phone thinks we're in the car park. The mail server won't talk to the network»[407]. Així, utilitzant el concepte de Dennett conegut com "postura intencional" (*the*

405 SHANAHAN, Murray (16/02/2023). "Talking About Large Language Models" en arXiv preprint arXiv:2212.03551, pàg. 1. Consultat el 8 d'abril de 2023 a: https://arxiv.org/pdf/2212.03551.pdf

406 *Ibídem*, pàg. 2

407 *Ibídem*, pàg. 3

intentional stance), Shanahan assumeix que en aquests casos aquest ús és completament inofensiu: «They are harmless because no-one takes them seriously enough to ask their watch to get it right next time, say, or to tell the mail server to try harder»[408]. No prendre-se'ls seriosament és la postura intencional pròpia d'aquell qui entén que l'ús d'aquests verbs (adonar-se'n, pensar o parlar: *realise*, *think*, *talk*) és figurat. El problema consisteix en prendre-se'ls seriosament, cosa que passa quan la postura intencional dels participants en la conversa és diferent entre ells o, simplement, és errònia, és a dir, interpreten aquell ús metafòric original en un sentit estricte i, en aquest cas, psicològic.

Per descartar objeccions, Shanahan considera els diferents escenaris en els quals es fan servir aquests LLM -a part de les consultes sobre informació del món (tipus Trivial)- com els models de conversió de text a imatge (DALL-E, VilBERT o Flamingo) o també en robòtica (SayCan), la seva especialitat. En aquests casos es podria al·legar que sí que es produeix certa relació amb la realitat, ja que, a diferència de la resposta a preguntes tipus Trivial, en el reconeixement i confecció d'imatges o en la robòtica hi ha uns sensors que capten elements del món. Tanmateix, la conclusió és la mateixa:

> A bare-bones LLM doesn't "really" know anything because all it does, at a fundamental level, is sequence prediction. Sometimes a predicted sequence takes the form of a proposition. But the special relationship propositional sequences have to truth is apparent only to the humans who are asking questions, or to those who provided the data the model was trained on. Sequences of words with a propositional form are not special to the model itself in the way they are to us. The model itself has no notion of truth or falsehood, properly speaking, because it lacks the means to exercise these concepts in anything like the way we do.[409]

Aquests *sequence prediction* té com a unitat bàsica el *token*, la unitat més petita del qual és el caràcter, però que també és escalable a parts de paraula o paraules senceres. Així, generar una frase consisteix en posar un *token* al costat d'un altre de forma que imiti, de la forma més creïble possible, com ho escriuria un humà. Per tant, la clau està en el grau de fiabilitat d'aquesta imitació (*mimicking human language* ho anomena). L'ús d'aquest terme, "imitar", ha estat bastament tractat en la filosofia, especialment des de que Plató el va utilitzar per descriure el que fan els artistes quan representen obres: imiten un personatge de forma que soni tan similar com seria possible. Plató diferencia aquesta imitació (*μίμησις - mīmēsis*) d'un altre tipus de semblança, la participació (*μέθεξις – méthexis*). La participació és allò que ens permet reconèixer una figura concreta com a exemplar d'una figura més general, per exemple, la pilota que xutem (de cuir, de color vermell i negre i un pel gastada) com un tipus de pilota entre totes les pilotes possibles (el prototip de pilota o

408 *Ídem.*

409 *Ibídem*, pàg. 5.

la idea de pilota). Sembla que la confusió, traduïda a un llenguatge platònic, es podria explicar de la següent forma: mentre que els especialistes en LLM entenen que l'ús dels verbs "pensar", "creure" o "saber" és estrictament imitatiu, els no especialistes els fan servir en un sentit participatiu.

Ara bé, perquè fos possible un ús participatiu d'aquests verbs, caldria que hi hagués una relació entre el *token* i la realitat, és a dir, que el *token* fos una unitat de significat en si mateixa, cosa que, tenint en compte que el *token* no deixa de ser una unitat d'un univers matemàtic, tal i com s'ha definit en el context dels LLM, seria sempre arbitrari. Aquesta demostració ja la va fer David Hume quan va plantejar el problema del concepte "d'igualtat i d'inigualat"[410]: un *token* només podria ser igual a una unitat amb significat si prèviament existís un patró sobre el que haver creat la relació entre aquest *token* i la realitat; tanmateix, aquest patró només es podria haver creat a partir d'un *token* anterior que ja es relacionés, amb significat (participativament), amb la realitat, i així infinitament. Per tant, qualsevol relació d'un *token* amb la realitat és gratuïta en la mesura que es basa en un patró concret confeccionat, en el cas del LLM, a partir de la comparació d'un elevadíssim nombre de textos. La relació d'aquells textos amb la realitat no ve donada en els mateixos textos, sinó que la donen el lectors quan interpreten que el terme "Lluna" en la frase "Qui va ser el primer home en trepitjar la Lluna", implica l'existència d'un satèl·lit que orbita al voltant de la Terra i d'un home, a qui coneixem pel nom d'Armstrong, que hi va clavar el peu. És a dir, el que no pot fer un LLM és entendre la participació del *token* "Lluna" respecte el satèl·lit que orbita la Terra, sinó que la pren com a mera imitació gràfica, sonora o visual.

Per tant, a l'igual que a la muntanya, hi ha dreceres que només els especialistes haurien de prendre en la mesura que en coneixen el seu sentit i els seus límits (i com no extraviar-s'hi), mentre que són perilloses per aquells que no entenen que la postura intencional pròpia davant d'aquests fenòmens és evitar de prendre-se'ls seriosament. Els LLM són bons imitadors dels textos produïts fins ara pels humans, funcionalitat que pot ser útil actualment, ja que hi ha moltes feines que consisteixen en tractament de textos, des de tasques administratives a comunicatives o comercials. De fet, és fàcil d'imaginar que, donada una professió, quan més elevada sigui la relació amb textos i aquests siguin de característiques similars, amb més èxit es podran utilitzar en aquests àmbits els LLM. Per la resta de casos, segueix sent millor prendre el camí més llarg.

Així acaba Shanahan: reconeixent que una alternativa seria introduir nous termes per referir-se als processos que imiten (o interpretem des de) fenòmens psicològics. Tanmateix, tal i com ell mateix admet: «it takes time for new language to settle, and for new ways of talking to find their

410 HUME, David (1740). *Abstract of a Book lately Published; Entitled, A Treatise of Human Nature, &c. Wherein the Chief Argument of that Book is farther Illustrated and Explained*, pàg. 9. Consultat el 9 d'abril de 2023 a: https://www.earlymoderntexts.com/assets/pdfs/hume1740.pdf

place in human affairs. It may require an extensive period of interacting with, of living with, these new kinds of artefact before we learn how best to talk about them».[411]

Requereix temps.

La computació i els límits del llenguatge

La relació entre les idees, les paraules i les coses és un dels temes clàssics de la filosofia. Al llarg de la seva història, ha aparegut sota diferents formes: en el *Cràtil* de Plató, Sòcrates juga amb Cràtil, que defensa la relació natural entre les paraules i les coses, i Hermògenes, que defensa que la relació és convencional. En el *Parmènides*, és Plató qui juga amb el lector en caricaturitzar el problema de la multiplicitat. El problema dels universals és la hipostatització cristiana d'aquest mateix joc, però ara pres seriosament i podent comportar la mort. A partir de la modernitat i especialment amb la voluntat divulgativa tan característica de la filosofia anglo-parlant, l'ús d'exemples mentals comença a ser una forma agradable i gràcil de traduir aquests problemes: el que en Kant són exemples extrems per fer entendre als seus alumnes quin és el tema i quin no[412], es converteixen en dilemes la sola presentació dels quals ja fixa la interpretació.[413]

Com qualsevol problema clàssic, independentment de la seva forma epocal, sempre acaba tornant a aparèixer, ja que denota un límit humà. Així, en el camp de la IA, els problemes clàssics reapareixen, però passats pel sedàs de la teoria de la informació, com explica Lyotard[414]. En aquest sentit, Erik J. Larson defensa que, mentre la IA clàssica (la simbòlica) prenia per innata tota possibilitat de coneixement i per això aplicava mètodes deductius, la IA generativa basada en algoritmes d'autoaprenentatge aplicats en *big data*, suposen una *tabula rasa* a la qual es va escrivint o moblant a partir de processos inductius[415]. Per altra banda, determinar què és una dada i què no és, és a dir, separar el soroll de les dades, amb la digitalització es converteix en un problema discret: només allò que es pot digitalitzar passa a ser emmagatzemat en estructures de bits; el que queda darrere, passa a ser oblidat, com recorda Weizenbaum[416]. Per això el context, que és allò que queda

411 SHANAHAN, Murray (16/02/2023). "Talking About Large Language Models" en *arXiv* preprint arXiv:2212.03551, pàg. 11. Consultat el 8 d'abril de 2023 a: https://arxiv.org/pdf/2212.03551.pdf

412 Un bon exemple d'això és la *Fonamentació de la metafísica dels costums*, i els exemples del botiguer, el suïcida o el gotós, entre d'altres.

413 En l'entorn la IA agrada tractar el dilema del cotxe autònom que ha de decidir si atropellar a un bebè o a una velleta, sense tenir en compte que els cotxes no decideixen (executen ordres) i que quan hi hagi cotxes realment autònoms (nivell 5), sempre podran sortir volant.

414 LYOTARD, Jean-François (1979). *La condición postmoderna*, Madrid, Cátedra, 2022.

415 LARSON, Erik J. (2021). *The myth of artificial intelligence: why computers can't think the way we do*, Londres, The Belknap Press of Harvard University Press, 2021, pàg. 4.

sempre darrere, però que permet entendre el que apareix davant, és el gran problema de la digitalització (*the frame problem*).

Tanmateix, els camins que han portat a una digitalització com la d'avui en dia, amb fitxers, ratolins i pantalles, hagués pogut anar per uns altres vorals, com explica Jaron Lanier arran de la imposició dels fitxers si Steve Jobs no els hagués introduït al Mac[417]. De fet, la computació mateixa no hauria de per què ser digital, si no fos perquè el silici és un semiconductor abundant[418]. En els orígens, una computadora era una persona que computava, és a dir, que feia càlculs[419], i el terme *analògic*, tot i l'ús específic que se'n fa en electrònica precisament per oposar-lo a allò digital (ús que no va entrar a la llengua anglesa fins la dècada de 1940[420]), etimològicament i encara en altres camps semàntics simplement indica que quelcom no és allò a què hom es vol referir exactament sinó que està *per sobre d'allò* i que posem *enlloc d'allò*, com equivalent a allò, i que prenem per allò a efectes pràctics: «The adjective *analog* comes from the word *analogy*. The concept behind this mode of computing is that instead of computing with numbers, one builds a physical model, or analog, of the system to be investigated»[421]. D'aquí l'ús d'analogia com a proporció o semblança en els camps matemàtics i lingüístics. També el terme *virtual* està perdent el seu significat original, com explicava Weizenbaum en una llarga entrevista el 2006: ««[…] "it was virtually night", that would have meant, even though it wasn't night it has all the characteristics that a person connects with the night»[422]. També quan es deia que un problema era *artificial*, volia dir que no era un problema natural, sinó que havia estat ideat simplement per poder servir per a un cert propòsit; un

416 Citat en WEIZENBAUM, Joseph (1976). *Computer Power and Human Reason*, Londres, Penguin Books Ltd, New Ed, 1984, pàgs. 237-238. Weizenbaum cita una anècdota de Philip Morris, professor de física al MIT, qui explicava com es va renunciar a digitalitzar tots els mapes sismològics d'abans de 1961, cosa que oblidava i, *de facto*, condemnava tota la informació que poguessin contenir.

417 LANIER, Jaron (2010). *You are not a gadget*, Londres, Penguin Books, 2011, pàgs. 12-13.

418 La quantitat de silici representa el 28% del mantell terrestre. Vid. SALVAT ROVIRA, N. (2021). "L'estructura bàsica de la vida" en *Biologia on-line*, Vol. 10, Núm. 1, febrer 2021. Recuperat el 3 d'agost de 2021 de: https://revistes.ub.edu/index.php/b_on/article/view/33883/33387

419 Computer (n). *Online Etymology Dictionary.* Consultat el 20 de juliol de 2024 a: https://www.etymonline.com/word/computer

420 MALL, A. (2003). "analog,digital" en *Theories of Media*, University of Chicago. Consultat el 4 d'agost de 2021 https://csmt.uchicago.edu/glossary2004/analogdigital.htm

421 CAMPBELL-KELLY, M. i ASPRAY, W. (1996). *A History of the Information Machine*, Nova York, Basic Books, 2013, pàg. 46.

422 WEIZENBAUM, Joseph; WENDT, Gunna (2006). *Islands in the Cyberstream. Seeking Havens of Reason in a Programmed Society*, Sacramento, Litwin Books, 2015, pàg. 120

propòsit que, d'alguna manera, es jutjava com a fals, poc natural[423]. Quan algú es plantejava un problema que pensava que era artificial, no buscava una solució (de fet, etiquetar-lo d'aquesta manera ja suposava algun tipus de desqualificació) i, en qualsevol cas, si no hi havia més remei que proposar una solució, no es pretenia que la solució rebuscada que se li havia donat fos extrapolable a qualsevol altre problema del mateix tipus (perquè no hi hauria d'haver problemes del mateix tipus: era artificial!)[424].

Per tant, d'una forma o d'una altra, se sabia des del principi: ja sigui per analògic, ja sigui per virtual o artificial, sempre ha quedat palès que allò només era un model, no pas la realitat, com les metàfores. Mentre que les metàfores juguen en els límits del llenguatge, per això són tan difícils de digitalitzar, tota la computació és en ella mateixa un gran metàfora. Per això, en la segona part d'aquest treball s'analitzarà l'origen de la metàfora computacional i també la seva inversió, condició necessària per confeccionar una etologia digital.

423 L'arrel del terme "artificial" és la mateixa que "artificiós".

424 Exemples reals d'això són les declaracions de Ricard Gomà el 15/06/2010 sobre un decret de l'alcalde de Barcelona d'aquell moment, Jordi Hereu, per prohibir el burca a les piscines municipals: «Prohibir el burca és generar un problema artificial» (Consultat el 20 de juliol de 2024 a: https://www.ccma.cat/tv3/alacarta/els-matins/prohibir-el-burca-es-generar-un-problema-artificial/video/2966870/). També en anglès: «An artificial solution to an artificial problem – Tax avoidance and the Dukeries case» (Consultat el 20 de juliol de 2024 a: https://www.wilberforce.co.uk/an-artificial-solution-to-an-artificial-problem-tax-avoidance-and-the-dukeries-case-by-f-moeran/).

Segona part

Claiming that in any way these machines are
"brains" is like claiming that walking
sticks are legs.

Piero Scaruffi, *Intelligence is not Artificial*

5. La metàfora computacional

Per entendre el paper de la metàfora computacional en la confecció d'una etologia digital, cal analitzar primer el paper de la metàfora en la ciència en general. En aquest treball, s'estudiaran tres propostes contemporànies i pioneres: la de Max Black, la de Richard Boyd i la de Thomas Kuhn. També es plantejaran algunes de les crítiques que es fa a l'ús de metàfores en la ciència, ja que són crítiques extrapolables també a l'ús de la metàfora computacional.

Seguidament, es farà una anàlisi de tres moments històrics de la metàfora computacional: des de la seva aparició com a projecte explícit i amb un ús no consolidat entre 1948 i 1957 (de Norbert Wiener a John von Neumann); els primers indicis de la seva consolidació, confirmats per una primera crítica de Jospeh Weizenbaum el 1972 i el reconeixement ple de la seva funcionalitat en el camp de la psicologia el 1979; i, finalment, un tercer moment de deconstrucció, per una banda, en tant que punt comú de partida el 1984 i, per l'altra, de separació o expressa diferenciació en el naixement d'una nova metàfora que deixa enrere la metàfora computacional i abraça la metàfora informàtica o del programador (que no deixa de ser la seva inversió), en un text de Pinker de 1997.

5.1 L'ús de la metàfora en la ciència

L'ús de la metàfora en la ciència ha estat àmpliament estudiat i una de les primeres obres de referència és la de Max Black (1909 - 1988), tant a l'article publicat el 1954 i titulat "Metaphor", com especialment a l'obra en la qual va desenvolupar més extensament la seva idea del paper de la metàfora en la ciència, *Models and Metaphors* (1962). Black parteix de la següent premissa irònica: «To draw attention to a philosopher's metaphors is to belittle him –like praising a logician for his beautiful handwriting»[425]. En l'estudi de les metàfores, Black posa de manifest que quan s'afirma "L'home és un llop" es produeix algun tipus d'interacció entre el significat tant d'home com de llop. No hi ha una mera substitució del contingut semàntic d'home pel de llop (com afirmaria la proposta coneguda com *substitution view of metaphor*), ni tampoc una comparació encoberta (*comparison view*) que seria possible explicitar en forma de símil (com sembla suggerir Aristòtil: «Metaphor

425 BLACK, Max (1954). "Metaphor" en *Proceedings of the Aristotelian Society*, New Series, Vol. 55 (1954), pàg. 273. Consultat el 30 de març de 2024 a: https://www.jstor.org/stable/4544549

consists in giving the thing a name that belongs to something else; the transference being either from genus to species, or from species to genus, or from species to species, or on grounds of analogy(l457b)»[426]). Per Black, alguns atributs contextuals d'home es troben, interactuen (per això l'anomena *interaction view*), amb alguns continguts contextuals de llop, en concret, quan es presenta els humans com a odiosos i alarmants, característiques que, segons Black, no s'atribueixen als llops en tot moment, sinó només als llops quan fan de llops en l'imaginari col·lectiu (cosa diferent al que tindria en ment un especialista en llops o un etòleg). En concret, per a Black, aquesta interacció afecta el significat dels dos membres: «If to call a man a wolf is to put him in a special light, we must not forget that the metaphor makes the wolf seem more human than he otherwise would»[427]. Black defensa que només les metàfores per interacció com aquestes tenen valor filosòfic, ara bé, que tinguin un valor filosòfic no els dona una legitimació científica. De fet, com observa Richard Boyd (1942-2021) en un article publicat el 1979 i titulat "*Metaphor and theory change: What is metaphor" a metaphor for?", Black està relegant a una segona posició el valor de la metàfora en la ciència:

> In particular, in this view [Black's view], one should expect that when metaphorical language is employed in a scientific context, its function should either lie in the pretheoretical (prescientific?) stages of the development of a discipline, or in the case of more mature sciences, it should lie in the realm of heuristics, pedagogy, or informal exegesis, rather than in the realm of the actual articulation or development of theories.[428]

Per tant, per una banda, en Black la metàfora té valor filosòfic en la mesura que no té valor científic (a excepció d'un paper inspiratori a nivell precientífic o un paper simplificador a nivell pedagògic). I, per altra banda, la proposta de Black conté un indici del que s'ha anomenat aquí la inversió de la metàfora computacional (IMC), ja que la *interaction view* pressuposa un intercanvi de valor entre el subjecte principal i el subjecte secundari (l'home i el llop respectivament en l'exemple). Ara bé, costa d'imaginar com aquesta metàfora fa veure els llops més humans. Si bé cal admetre que els atributs associats al llop simplifiquen realment el que és un llop (i més encara des del punt de vista d'un especialista en llops), la transferència de significat no és de doble sentit, ja

426 ARISTÒTIL, *Poètica*, 1457b. Citat en BLACK, Max (1954). "Metaphor" en *Proceedings of the Aristotelian Society*, New Series, Vol. 55 (1954), pàg. 284. Consultat el 30 de març de 2024 a: https://www.jstor.org/stable/4544549

427 BLACK, Max (1954). "Metaphor" en *Proceedings of the Aristotelian Society*, New Series, Vol. 55 (1954), pàg. 291. Consultat el 30 de març de 2024 a: https://www.jstor.org/stable/4544549

428 BOYD, Richard N. (1979). "*Metaphor and theory change: What is metaphor" a metaphor for?" en en *Metaphor and thought,* Editor Ortony A, Cambridge, MA, Cambridge University Press, 1979, pàg.482.

que no hi ha cap atribut d'humà que acabi sent heretat pel llop, com proposa Black. Ara bé, si s'inverteix l'ordre de la metàfora afirmant que "El llop és un home", en aquest cas sí que s'estaria predicant que el llop posseeix algun atribuït associat habitualment a l'home (sociabilitat?). Per tant, en les metàfores es produeix aquesta transmissió de significat, però només en un sentit: del segon membre cap al primer membre. I no d'aquells atributs que un especialista en el segon membre podria definir, sinó d'alguns atributs concrets i que, normalment, són els que circulen per l'imaginari col·lectiu. Per això, les metàfores són contextuals.

Boyd retreu a Black aquesta manca de valor científic que atribueix a les metàfores: «There exists an important class of metaphors which play a role in the development and articulation of theories in relatively mature sciences. Their function is a sort of *catachresis* - that is, they are used to introduce theoretical terminology where none previously existed»[429]. Boyd defensa que en les metàfores que fan aquesta funció en la ciència és sempre difícil acabar d'especificar quina similitud o analogia hi ha entre els dos membres de la comparació, ja que la seva gràcia rau en la seva capacitat de mantenir obert el problema (*open-ended*). És això el que els dona valor i potència: així, segons Boyd, la gràcia d'anomenar-los "forats de cuc" (*worm-holes)* o descriure els àtoms com "sistemes solars en miniatura" és que mantenen viva la investigació científica i donen joc per introduir nou vocabulari[430]. Tanmateix, sembla que Boyd no considera que aquesta obertura també es pot enquistar i que en cert moment, si més no lluny d'un entorn especialitzat, pot ser igual d'obvi que els forats negres siguin forats (i més per a la representació que se n'ha fet en les pel·lícules de ciència ficció) que una boca de reg sigui una boca o les potes d'una taula unes potes. És a dir, hi ha un moment que l'obertura es tanca i aquestes metàfores queden fixades en el vocabulari, moment també en què perden la seva utilitat científica (tot i que segurament mantenen la seva funció pedagògica, com el cas de les representacions d'àtoms amb porexpan als instituts, encara que costi saber si estan estudiant astronomia o la constitució de la matèria).

Boyd manté que mentre hi ha obertura és perquè el problema està viu i que això reflecteix una forceig entre el llenguatge científic i el desordre i complexitat del món, cosa que confereix a les metàfores la següent funció: «More precisely, what I shall argue is that the use of metaphor is one of many devices available to the scientific community to accomplish the task of accommodation of language to the causal structure of the world»[431]. Per tant, aquestes imatges que traslladen les metàfores són com una peça d'acoblament o juntura (*joints*) que ajuden a fer encaixar un cos teòric en el fenomen que pretenen representar: «What I shall argue here is that the employment of

429 *Ídem.* La cursives és de Boyd.

430 *Ídem.*

431 *Ibídem*, pàg. 483.

metaphor serves as a nondefinitional mode of reference fixing which is especially well suited to the introduction of terms referring to kinds whose real essences consist of complex relational properties, rather than features of internal constitution»[432]. Per això, per a Boyd, entendre bé aquestes metàfores científiques podria facilitar la comprensió de la naturalesa de les essències reals. En aquest sentit, sembla que des de la proposta de Boyd, les metàfores no es limitarien a tenir un paper inspiratori, sinó que permetrien una relació directe entre les idees i la realitat quan encara manca un vocabulari més específic (i s'entén que definicional d'aquesta realitat). Boyd exemplificarà això amb la metàfora computacional, com es veurà en el següent apartat.

Thomas Kuhn (1922-1996) discrepa concretament d'aquest últim punt, a saber, que les metàfores puguin predicar res de la realitat, bàsicament perquè la ciència no consisteix en un mer conjunt d'enunciats certs sobre la realitat. En un article titulat "Metaphor in science", també de 1979, Kuhn comença caracteritzant la postura de Boyd en aquest darrer article, però no només pel que diu, sinó pel que implica allò que diu: «If Boyd is right that nature has "joints" which natural-kind terms aim to locate, then metaphor reminds us that another language might have located different joints, cut up the world in another way»[433]. Si per Boyd aquestes metàfores feien de juntures entre les idees i la realitat, Kuhn posa l'èmfasi en què precisament, en tant que metàfores, poden acoblar-se de diferents maneres i per diferents llocs, ja que no s'estan acoblant a la realitat, sinó al model de la realitat que utilitza la teoria: «Without its aid, one cannot even today write down the Schrodinger equation for a complex atom or molecule, for it is to the model, not directly to nature, that the various terms in that equation refer»[434]. Aquesta referència al model com a constructe intermedi entre les idees i al realitat, i les metàfores com a juntura per relacionar les idees amb els models a través de paraules, deixa en un segon pla inassolible la realitat i condensa la tasca de la ciència a la construcció de models útils i funcionals:

> I shall close with a metaphor of my own. Boyd's world with its joints seems to me, like Kant's "things in themselves," in principle unknowable. The view toward which I grope would also be Kantian but without "things in themselves" and with categories of the mind which could change with time as the accommodation of language and experience proceeded. A view of that sort need not, I think, make the world less real.[435]

---

432 *Ídem*.

433 KUHN, Thomas S. (1979). "Metaphor in Science" en *Metaphor and thought,* Editor Ortony A, Cambridge, (MA), Cambridge University Press, pàg. 537. Consultat el 10 de juliol de 2024 a: https://www.cambridge.org/core/books/abs/metaphor-and-thought/metaphor-in-science/291E8A7ADF9A427260C5C6C8653A1F1F

434 *Ibídem*, pàg. 538.

435 *Ibídem*, pàg. 542.

La constitució d'aquests models i el llenguatge formal pel qual es construeixen són una de les peces centrals per entendre la inversió de la metàfora computacional, com es veurà més endavant.

Per tant, mentre la proposta de Black atorga a les metàfores un paper inspirador i pedagògic en la ciència, Boyd les fa servir com a juntures que relacionen les idees i el món a partir d'un isomorfisme estàtic, mentre que Kuhn, sense trencar aquest isomorfisme, el situa entre les idees i les paraules, ja sigui en tant que metàfores (llenguatge natural) ja sigui en tant que models (llenguatge formal), i exclou un accés directe i definitiu al món.

Ara bé, l'ús de la metàfora en la ciència també ha tingut detractors, com recopilen Cynthia Taylor i Brayan M. Dewsbury en un article titulat "On the Problem and Promise of Metaphor Use in Science and Science Communication" (2018)[436]. La crítica principal és que les metàfores, mentre il·luminen una part de l'analogia, en deixen a les fosques altres. Posen diversos exemples de metàfores que distorsionen el coneixement científic actual com aquelles que es fan en genètica: «If genes really do function as blueprints, we should expect a one-to-one correspondence between particular genetic "instructions" and phenotypic out-comes in organisms, with limited input from the environment in structuring variation between individuals. Yet this is not the case»[437]. I també qüestionen com una visió bel·licista del món penetra en la descripció científica al parlar d'espècies invasores, agents infecciosos o cèl·lules segrestades per virus. Per altra banda, Massimo Pigliucci i Maarten Boudry, en un article de 2010 titulat "Why Machine-Information Metaphors are Bad for Science and Science Education", rastregen fins a David Hume una de les primeres crítiques contra l'analogia mecanicista, i citen el moment en que Philo, el personatge que proposa recuperar la metàfora organicista clàssica, assumeix que qualsevol metàfora planteja limitacions: «[I]n such questions as the present, a hundred contradictory views may preserve a kind of imperfect analogy, and invention has here the full scope to exert itself»[438]. De fet, Pigliucci i Boudry també assenyalen el perill de caure en una defensa del disseny intel·ligent derivada de l'associació entre la naturalesa i el rellotger de l'univers, estructura conceptual que ja s'ha detectat en alguns plantejaments dels autors estudiats en la primera part d'aquest treball. En aquest sentit, citen un advertiment de Charles Darwin: «Biology is the study of complicated things that give the appearance of having been

436 TAYLOR, Cynthia; DEWSBURY, Brryan M. (2018). "On the Problem and Promise of Metaphor Use in Science and Science Communication" en *Journal of Microbiology & Biology Education*, 19(1), 2018. Consultat el 23 de juliol de 2024 a: https://doi.org/10.1128/jmbe.v19i1.1538

437 *Ibídem*, pàg. 3

438 HUME, David (1779). *Dialogues concerning natural religion*, (2nd ed), Hackett, pàg. 49. Citat en PIGLIUCCI, Massimo; BOUDRY, Maarten (11.06.2010). "Why Machine-Information Metaphors are Bat for Science and Science Education" en *Sci & Educ* 20, 2011, pàg. 458. Consultat el 23 de juliol de 2024 a: https://doi.org/10.1007/s11191-010-9267-6

designed for a purpose»[439]. Darwin s'estava referint explícitament a la proposta de William Paley, creador de l'argument del rellotger.

En qualsevol cas, tant defensors com detractors de l'ús de la metàfores en ciència accepten, no només la tesi de Lakoff i Johnson, és a dir, que les metàfores condicionen la forma de veure el món de la vida quotidiana (pretematitzada, amb biaixos incorporats), sinó que amb Theodore L. Brown, també condicionen la forma de fer ciència (tematitzada, pretesament sense biaixos), en la mesura que impregnen tant models com hipòtesis i teories científiques: «The fact that metaphor is so inextricably a part of the fabric of science also means that it plays many roles. Scientists use metaphorical reasoning to interpret observational data, creating models to account for new observations and to reinterpret older data»[440]. En aquest sentit, sembla que hi hauria d'haver consens amb la frase d'obertura d'aquest treball: «The price of metaphor is eternal vigilance»[441]. Tanmateix, en el següent apartat s'analitzarà per què en el cas concret de la metàfora computacional això no és així.

5.2 La metàfora computacional: una perspectiva històrica

Els primers tantejos amb la metàfora computacional que, durant l'elaboració d'aquest treball, s'han trobat són els de Norbert Wiener de 1948. Juntament amb ell i durant una desena d'anys, s'escriuen una sèrie de textos que ajudaran a assentar la metàfora computacional, no com una mera hipòtesi, sinó posteriorment com un axioma. Tots aquests textos inicials, que s'analitzaran en el següent apartat, són textos especulatius en els quals cal fer un esforç imaginatiu i justificatiu per donar sentit a la comparació entre l'ordinador i el cervell, normalment en aquest ordre. El primer text és de 1948 i l'últim d'aquest període de construcció, de 1957. Després d'aquests dos textos, ja s'ha trobat, el 1972, la primera crítica explícita contra l'ús de la metàfora computacional: un article de Weizenbaum en el qual encara la crítica mateixa té un caràcter de provisionalitat. Per tant, entre 1957 i 1972, la metàfora s'ha assentat suficientment, si més no en l'entorn acadèmic, com per poder-se començar a qüestionar.

---

439 DARWIN, Charles (1859). *On the origin of specie*, http://www.darwin-online.org.uk/. Citat en PIGLIUCCI, Massimo; BOUDRY, Maarten (11.06.2010). "Why Machine-Information Metaphors are Bat for Science and Science Education" en *Sci & Educ* 20, 2011, pàg. 458. Consultat el 23 de juliol de 2024 a: https://doi.org/10.1007/s11191-010-9267-6

440 BROWN, Theodore L. (2008). *Making Truth: Metaphor in Science*, Illinois, University of Illinois Press, 2008, pàgs. 184-185.

441 ROSENBLUETH, Arturo; WIENER,Norbert. Citat en PIGLIUCCI, Massimo; BOUDRY, Maarten (11.06.2010). "Why Machine-Information Metaphors are Bat for Science and Science Education" en *Sci & Educ* 20, 453–471 (2011), pàg. 454. Consultat el 23 de juliol de 2024 a: https://doi.org/10.1007/s11191-010-9267-6

Ara bé, tampoc és cap sorpresa que aparegui aquesta analogia en aquest moment de la història: després de la proposta mecanicista moderna, l'analogia entre el cos i una màquina s'ha anat actualitzant, com s'ha vist amb Zarkadakis, amb l'arribada de l'electricitat, el telègraf i, finalment, l'ordinador. Per tant, la versió de l'autòmat del segle XVIII és el robot del segle XX, cosa que porta a què l'ordinador sigui el cervell d'aquest robot. Així d'evident ho veu Nil Nilsson, un dels fundadors del camp de la IA:

> Against this background of prediction successes and failures, I hesitate to make any that do not seem rather obvious. Except, I will predict that someday we'll have human-made artifacts with levels of intelligence (in all of its manifestations) equalling and exceeding that of humans. I make that prediction because I believe that we humans are machines (for what else could we be?) and that eventually we'll be able to build machines that can do whatever we can do because there will be economic as well as scientific reasons for doing so.[442]

Aquesta interrogació retòrica (*what else could we be?*) estava molt estesa entre els fundadors de la IA, i també hi és en Shannon, pare de la teoria de la informació, el 1977: «"You bet," he replies, when asked whether he thinks machines can think. "I'm a machine and you're a machine, and we both think, don't we?"»[443]. De fet, cal recordar que Warren McCulloch i Walter Pitts ja havien proposat, el 1943, entendre la neurona com una unitat lògica ("all-or-none" anomenen al seu comportament binari) que operava a través d'*inputs* i *outputs*, termes que ells encara no fan servir, sinó que són introduïts, aquell mateix any, per Arturo Rosenblueth, Norbert Wiener i Julian Bigelow en un article titulat "Behavior, Purpose and Teleology". Fou en aquest text on per primera vegada s'utilitzen fora del context comptable i electrònic en el qual eren habituals (el significat comptable de *input* és equivalent al d'imputar, com, per exemple, imputar una despesa, i es recull aquest ús des de 1753; en canvi, el significat electrònic és molt més recent, de 1902, i es feia servir per indicar l'energia proporcionada a una aparell o a una màquina[444]):

> Given any object, relatively abstracted from its surroundings for study, the behavioristic approach consists in the examination of the output of the object and of the relations of this output to the input. By output is meant any change produced in the surroundings by the object.

442 NILSSON, Nils J. (2010). *The Quest for Artificial Intelligence. A History of Ideas and Achievements*, New York, Cambridge University Press, 2010, pàg. 515.

443 SHANNON, Claude E. (1992). *Collected Papers*, Ed. N.J.A. Sloane i Aaron D. Wyner, Nova York, IEEE i Wiley Interscience, 1992, pàg. XVII.

444 input (n). *A Online Etymology Dictionary*. Consultat el 15 d'agost de 2024 a: https://www.etymonline.com/word/input

> By input, conversely, is meant any event external to the object that modifies this object in any manner.[445]

I no és fins 1946 que són descrits com a òrgans externs a un ordinador per Arthur W. Burks, Herman H. Goldstein i John von Neumann en un article titulat "Preliminary Discussion of the Logical Design of an Electronic Computing Instrument". Tres anys més tard, el 1949, Donald O. Hebb, va defensar que les neurones eren les unitats bàsiques del pensament.[446]

Per tant, en la mesura que les humans han de ser màquines (*what else*), el cervell és també un tipus de màquina, és a dir, una computadora, i que ja hi ha una equivalència clara entre la unitat bàsica del cervell (la neurona) i la unitat bàsica de l'ordinador (el bit), semblen faves comptades desentranyar l'operativa isomòrfica entre unes i altres, cosa que permetria dissenyar una màquina que pogués pensar. És en aquest context de si una màquina pot pensar que Searle, als anys 80, partint de l'acceptació de la premissa que el cervell és una màquina, defensa que la ment no és el *software* del cervell: «But the equation, "mind is to brain as program is to hardware" breaks down at several points among them the following three»[447] (i els tres punts són, resumidament, que simular no és duplicar; memoritzar no és aprendre; i que així com els estats mentals són intencionals, els llenguatge formals, no). I és en aquest mateix article que Searle fa una diferència clàssica entre IA forta i IA dèbil: mentre que el projecte de la IA dèbil és el projecte d'aconseguir utilitzar l'ordinador com una eina molt potent en l'estudi de la ment, la IA forta és aconseguir duplicar la ment, per tal de construir una màquina que realment pensi[448]. I Dennett, també a la mateixa dècada, està disposat a acceptar que una màquina pensa sempre que superi honestament el test de Turing (precisament, és quan posa l'exemple de la gran ciutat): «My philosophical conclusion in this paper is that any computer that actually passed the Turing test would be a thinking thing in every theoretically interesting sense».[449]

---

445 ROSENBLUETH, Arturo; WIENER,Norbert; BIGELOW, Julian (1943). "Behavior, Purpose and Teleology" en *Philosophy of Science*, Vol. 10, No. 1, gener, 1943, pàg. 18.

446 NILSSON, Nils J. (2010). *The Quest for Artificial Intelligence. A History of Ideas and Achievements*, New York, Cambridge University Press, 2010, pàg. 17.

447 SEARLE, John R. (1980). "Mind, brains, and programs" en *Behavioral and Brain Sciences*, Cambridge, Cambridge University Press, 3 (3), pàg. 441. Consultat el 22 de juliol de 2024 a: https://www.cambridge.org/core/journals/behavioral-and-brain-sciences/article/abs/minds-brains-and-programs/DC644B47A4299C637C89772FACC2706A

448 *Ibídem*, pàg. 417.

449 DENNETT, Danniel (1985). "Can Machines Think" en *How We Know*, Ed. Michael Safto, San Francisco, Harper & Row, 1985, pàg. 140. Consultat el 22 de juliol de 2024 a: https://www.researchgate.net/publication/285475907

Per tant, tot científic, per tant, materialista per alguna banda (això inclou els dualistes), que accepti l'analogia entre cos i màquina (o mecànica), ha d'acceptar també l'analogia entre cervell i computació, en la mesura que la tesi Church-Turing explica com es pot computar a partir d'elements mecànics. Es poden presentar dos tipus de discrepàncies dins d'aquesta visió: per una banda, el tipus de discrepàncies que es fixen en les limitacions del concepte de *computació*. Per exemple, es pot qüestionar si tot enunciat és computable (i el propi Turing admet que no en "On computable numbers, with an application to the Entscheidungsproblem" (1936), doncs hi ha, com a mínim, el problema de l'aturada (*halting problem*); a part, també hi ha el problema dels adverbis i la seva dificultat de fixació quantitativa (per fixar un *lentament*, cal establir alguna relació entre la percepció d'un subjecte que fa una acció i el nombre d'unitats de temps que la mesuren, i això no només pot canviar per cada subjecte, sinó que també pot canviar per cada ocasió en què fa aquesta acció); després hi ha les accions que no es poden ni dir i, com defensa Weizenbaum, aquestes no es podran computar; i, finalment, també la idea que qualsevol matematització, i la computació és un tipus de matematització, configura un model, però un model no és la realitat, com diu Cathy O'Neil: «[...] todo modelo es, por su propia naturaleza, una simplificación. Ningún modelo puede incluir toda la complejidad del mundo ni los matices de la comunicación humana. Es inevitable que parte de la información importante se quede fuera»[450]. Per tant, no és evident que computar sigui equivalent a pensar, com es veurà també més endavant.

Per altra banda, hi ha el tipus de discrepàncies que es fixen en els límits del concepte *cervell*. És obvi que el coneixement que es té del cervell continua sent bastant inicial, tot i els avenços dels últims anys en neurociència –Javier de Felipe, professor d'Investigació en l'Instituto Cajal (CSIC) especialitzat en l'estudi microanatòmic del cervell, és així de rotund: «Pero otro tema distinto es crear un cerebro artificial, cuyo principal escollo, como veremos a continuación, no es solo la complejidad extraordinaria del sistema nervioso, sino nuestro desconocimiento sobre su organización estructural y funcional»[451].

També han aparegut discrepàncies en si es pot, dins d'aquesta visió mínimament materialista, diferenciar entre cervell i ment. Per von Neumann, el plantejament manca tant de sentit que la paraula *ment* només s'utilitza una sola vegada precisament per dir el que no se sap: «We are as ignorant of its nature and position as were the Greeks, who suspected the location of the mind in the

450 O'NEIL, Cathy (2017). *Armas de destrucción matemática. Cómo el Big Data aumenta la desigualdad y amenaza la democracia*, Capitán Swing Libros, Madrid, 2017, pàg. 30.

451 DE FELIPE, Javier (2022). *De Laetoli a la Luna: El insólito viaje del cerebro humano*, Barcelona, Crítica, 2022, pàg. 198.

diaphragm»[452]. Per tant, un dels responsables de la tematització inicial de l'analogia busca com trobar una equivalència entre un ordinador i el cervell, en cap cas la ment (segurament perquè dona per fet que la ment és un dels noms del cervell, com se suposa que va dir Minsky[453]). Ara bé, des de que Searle s'enfronta al problema de la intencionalitat (i també el de la computabilitat) a través de l'experiment mental de l'habitació xinesa, per tal de poder separar clarament els límits d'una sintaxis i d'una semàntica (cosa que li permet defensar que l'ordinador computa sintaxis, però desconeix, *de iure*, una semàntica), li acaba sorgint, per la porta del darrere, el problema del subjecte d'aquesta semàntica. Com és obvi, el problema és un clàssic de la filosofia traduït ara al paradigma del segle XX, per tant, s'adapta la relació cos-ànima a cervell-ment.

Un dels grans defensors que la ment és el software del cervell és Ned Block qui, en un exemple d'inversió de la metàfora computacional i des d'una perspectiva estrictament materialista, defensa que es pot anomenar ment al moviment de símbols que fa el cervell, interpretant així la ment com el *software* del cervell (*hardware*)[454].

En qualsevol cas, entre qui defensa explícitament l'isomorfisme perfecte entre cervell-ment (com Block o Johnson-Laird), aquells que utilitzen una paraula enlloc de l'altra i certa confusió afegida pel component ideològic determinat en cada posicionament, tot plegat afecta a diferents interpretacions de la metàfora computacional, la més evident de les quals és sobre l'aplicació poc clara del vocabulari de la ment (i.e., *reconèixer*, *pensar*, *imaginar* o *entendre*, però també *descriure*, *sumar, modular* o *computar*) en el vocabulari del cervell (i.e., *neurona*, *neurotransmissor*, *cèl·lules glials* o *anandamida*) i d'aquí al vocabulari de la informàtica (i.e., *bit*, *algoritme*, *input/output* o *dispositiu*). Sembla que la possible arrel comú d'aquesta aplicació seria la teoria de la informació de Shannon, que vincula el codi informàtic amb el codi genètic i el concepte d'entrada i sortida amb el d'emissor i receptor, entre d'altres.

---

452 NEUMANN, John von (1953). *The Computer & the Brain,* New Haven i Londres, Yale University Press, 2012, pàg. 62. Consultat el 22 de juliol de 2024 a: https://ia600707.us.archive.org/3/items/0300181116TheComputerBrain_201901/0300181116_The%20Computer%20Brain.pdf

453 Segons Nilsson, se suposa que Minsky va dir «The mind is a meat machine», mentre que Weizenbaum afirma que va ser Simons qui es va preguntar si «Is the brain merely a meat machine?». Ara bé, el fet que l'anècdota divergeixi també mostra amb quina facilitat s'intercanvien aquests dos termes dins del camp de la IA, que és el que aquí s'està defensant.

454 BLOCK, Ned (1990). "The mind as the software of the brain" en *An Invitation to Cognitive Science: Visual cognition*, Eds. Daniel N. Osherson & Edward E. Smith, Cambridge (MA), MIT Press, 2, 1990, pàgs. 377-425.

5.2.1 La construcció de la metàfora

Els primers tantejos sobre la metàfora computacional que, durant l'elaboració d'aquest treball, s'han trobat són els de Norbert Wiener (1894-1964), els Warren McCulloch (1898-1969), els de Claude Shannon (1916-2001) i els de John von Neumann (1903-1957). Wiener fa la seva primera proposta el 1948; McCulloch, el 1949; Shannon, el 1953; i von Neumann, el 1957. La influència entre aquests autors és evident, no només perquè participen junts a diferents congressos, sinó perquè coneixen els seus respectius treballs i, en alguns casos, en parlen abans de les publicacions oficials dels mateixos, com explica Steve J. Heims en una obra que descriu la relació entre von Neumann i Wiener, *John von Neumann and Norbert Wiener: From Mathematics to the Technologies of Life and Death* (1980): una biografia comparada d'ambdós matemàtics amb un capítol central, "A mutual interest, but…", en el qual es recompten les diferents trobades i la correspondència que van compartir (sembla que l'interès era mutu, però no la metodologia).

En qualsevol cas, hi ha un element compartit en tots ells: la metàfora no és quelcom des d'on es parteix, sinó una fita a la qual s'arriba amb esforç i justificació.

La primera intuïció: Norbert Wiener i la cibernètica (1948)

Wiener publica el 1948 un llibre, *Cybernetics: Or the Control and Communication in the Animal and the Machine*, que s'ha considerat l'obra fundacional, no només de la cibernètica tal i com s'entén actualment, sinó també del projecte IA. En aquesta, Wiener s'aproxima al cervell a partir de l'estudi comparatiu de la retroalimentació (*feedback*) en màquines i animals, en concret, entre el timó dels vaixells i les mans d'una persona amb tremolors[455]. En aquest context, Wiener proposa el terme *kybernetike* (cibernètica) en un sentit estricte, és a dir, el de timoner, i l'acció descriu com, depenent d'allò amb què topen, un timó o una mà canvien el seu moviment. En aquest sentit, la idea ja recull, embrionàriament, la necessària relació entre l'exterior i l'interior d'un sistema, és a dir, entre el món i jo, o, en vocabulari computacional, entre un sensor i el *software*. L'obra té una estructura aparentment poc sistemàtica i la comparació de l'ordinador i el cervell només és el tema central d'un capítol, "Computing Machines and the Nervous System" i, secundàriament, es torna a mencionar en el penúltim capítol de l'edició de 1948, "Cybernetics and Psychopathology" (la segona edició, la de 1961, incorpora dos nous capítols, però ja no mereixen l'atenció al no ser pioners).

Tanmateix i de forma contraintuïtiva, la primera menció d'alguna relació entre algun element humà i algun element mecànic no es dona en el context de la retroalimentació ni tampoc és del

455 NILSSON, Nils J. (2010). *The Quest for Artificial Intelligence. A History of Ideas and Achievements*, New York, Cambridge University Press, 2010, pàg. 27.

cervell amb l'ordinador, sinó en el context de l'autonomia i quan Wiener es planteja com hauria d'operar un ordinador: «The ideal computing machine must then have all its data inserted at the beginning, and must be as free as possible from human interference to the very end»[456]. Per tant, la primera idea és que en la computació, quanta menys intervenció humana hi hagi, millor. Aquesta obsessió amb l'autonomia serà una de les constants del projecte IA, en la mesura que s'interpreta que és com actuen els humans: davant d'un estímul, no esperen ordres exteriors per actuar. Ara bé, Wiener encara no és ben bé en aquesta versió (bàsicament perquè al projecte IA encara li falten uns 10 anys per inaugurar-se oficialment a Dartmouth), i encara fa servir un concepte d'autonomia que descriu millor un rellotge, el qual un cop has donat corda (o pila), no requereix de la intervenció de res ni ningú per operar perfectament fins que aquesta bateria s'esgota –és Weizenbaum qui assenyalarà aquesta analogia entre l'autonomia del rellotge i la de l'ordinador en un article de 1972 que s'analitzarà més endavant.

Seguidament, Wiener explica en quin sentit es podria fer una comparació entre un ordinador i el cervell: «It is a noteworthy fact that the human and animal nervous systems, which are known to be capable of the work of a computation system, contain elements which are ideally suited to act as relays. These elements are the so-called neurons or nerve cells»[457]. A part de comptar amb neurones que pot considerar-se que computen binàriament, el sistema nerviós també emmagatzema informació, per tant, té una memòria: «A very important function of the nervous system, and, as we have said, a function equally in demand for computing machines, is that of memory, the ability to preserve the results of past operations for use in the future»[458]. És en aquesta anàlisi de la memòria que Wiener posa de manifest com la memòria humana no és tan volàtil com la dels ordinadors: el cervell s'inicia una vegada i fins la mort d'una persona va guardant records, mentre un ordinador o un programa, cada vegada que es reinicia, hi ha una part de la memòria (la temporal) que s'esborra o variables globals que prenen el seu valor inicial: «Thus the brain, under normal circumstances, is not the complete analogue of the computing machine but rather the analogue of a single run on such a machine»[459]. Aquest tipus de reflexions denoten que la comparació encara és tendra, que s'està palpant per on pot anar, sense tenir molt clar com pot acabar (l'argument té un punt d'absurd: si no reinicies l'ordinador ni cap dels programes instanciats, l'analogia es recupera).

456 WIENER, Norbert (1948). *Cybernetics: Or the Control and Communication in the Animal and the Machine*, Cambridge (MA), MIT Press, 1985, pàg. 118

457 *Ibídem*, pàg. 120.

458 *Ibídem*, pàg. 121.

459 *Ídem.*

En aquest sentit, la base de la comparació encara és necessari que sigui especificada i no es pot donar per sabuda: «We have already spoken of the computing machine, and consequently the brain, as a logical machine. It is by no means trivial to consider the light cast on logic by such machines, both natural and artificial»[460]. La comparació és possible només en la mesura que es pot operar lògicament des d'ambdós sistemes, però, al mateix temps, en poder-se operar lògicament des d'ambdós sistemes, caldrà analitzar les limitacions i imperfeccions d'un i altre i veure quines són més fàcils de resoldre: «According to this, the study of logic must reduce to the study of the logical machine, whether nervous or mechanical, with all its non-removable limitations and imperfections»[461]. Aparentment, semblaria que aquesta afirmació posa en igualtat de condicions el cervell i l'ordinador, però ja ha dit anteriorment que en un ordinador, quanta menys intervenció humana hi hagi, millor i, després d'aquesta afirmació, Wiener ho aclareix:

> Psychology contains much that is foreign to logic, but—and this is the important fact—any logic which means anything to us can contain nothing which the human mind—and hence the human nervous system—is unable to encompass. All logic is limited by the limitations of the human mind when it is engaged in that activity known as logical thinking.[462]

Per tant, en tant que màquines lògiques, es pot fer una comparació entre l'ordinador i el cervell, però com que del que es tracta és d'operacions lògiques, hi ha una de les dues que està dissenyada específicament per computar, sense limitacions humanes. Això és el naixement de la cibernètica, no ja com a timoner que fa cops de pal a cegues per respondre a la corrent, sinó com a integració entre organismes vius i màquines.

Les últimes línies del capítol, tanmateix, semblen recuperar una mica de la perspectiva: a l'any 1948 és difícil que algú, davant d'una màquina que ocupa tota una habitació, que funciona amb relés o tubs de buit i que fa tant soroll com una fàbrica tèxtil, pugui trobar raonable un plantejament transhumanista (i el concepte de cíborg no apareixeria fins 13 anys més tard). Per tant, Wiener tanca la comparació sense acabar d'assentar la metàfora computacional: «The mechanical brain does not secrete thought "as the liver does bile," as the earlier materialists claimed, nor does it put it out in the form of energy, as the muscle puts out its activity. Information is information, not matter or energy. No materialism which does not admit this can survive at the present day»[463]. Ara bé, com a projecte experimental li segueix veient una sortida: «Nevertheless, the realization that the brain and the computing machine have much in common may suggest new and valid approaches to

460 *Ibídem*, pàgs. 124-125.

461 *Ibídem*, pàg. 125.

462 *Ídem.*

463 *Ibídem*, pàg. 144.

psychopathology and even to psychiatrics»[464]. De fet, bona part dels projectes informàtics que actualment s'utilitzen massivament es van crear, originalment, amb un objectiu pal·liatiu per algun col·lectiu desafavorit, com els lectors de textos automàtics per ajudar a persones amb dislèxia o els descriptors d'imatges per a cecs.

En conclusió, en aquest text, el primer en el qual s'ha trobat un intent de construcció de la metàfora computacional, la proposta és això, una proposta: no està consolidada i la metàfora encara no funciona per si sola, en part, perquè tampoc hi ha un imaginari col·lectiu que tingui molt ben definit cap dels dos termes de la comparació, ni l'ordinador ni el cervell. Per això, com també es veurà en les següents propostes, cal fer un esforç pedagògic per explicar bé com funciona tant un com l'altre.

El cervell com una màquina de computar: McCulloch (1949)

Un dels articles que més ha passat desapercebut, però que també col·labora, encara que sigui indirectament, en la confecció de la metàfora computacional és el que va publicar Warren S. McCulloch l'any 1949 titulat "The brain as a computing machine". Tot i ser un autor a qui es reconeix la paternitat de l'equivalència entre neurona i bit a l'afirmar que la neurona és una unitat lògica (article molt citat, publicat el 1943 juntament amb Walter Pitt i titulat "A logical calculus of the ideas immanent in nervous activity"), aquest article de 1949 és pràcticament ignorat en el seu moment: només té 10 citacions (i una d'aquestes citacions és del mateix McCulloch en un article posterior) abans de 1958, moment en el qual es publica l'obra de referència de von Neumann sobre el tema; tanmateix, una de les citacions la fa Claude Shannon l'any 1953, en un article titulat "Computers and Automata".

Es plantegen dues hipòtesis de per què l'article és tan poc reconegut: la primera té a veure amb la mateixa metàfora computacional, ja que tot i que durant l'article McCulloch fa un esforç per trobar similituds entre el cervell i una màquina de computar, el vocabulari que utilitza i, en part, l'imaginari des del qual treballa, és més propi de la metàfora electrònica que no pas de la computacional (entenent aquest segon terme des d'una perspectiva com l'actual, és a dir, digital). Ho mostren, per exemple, les següents frases inicials del text:

> Electrical engineers distinguish between problems of strong currents—or power engineering—and of weak currents or communication engineering. Computing machines, including brains, belong to the latter specialty. Man's brain is much the most complicated of computing machines

464 *Ídem.*

> and it requires power to keep its relays in the operating range of voltage. It is battery-operated, each relay having its own battery.[465]

L'especialista de referència és l'enginyer electrònic i el problema es relaciona amb relés, voltatges i bateries. És cert que s'afirma que el cervell és un tipus de màquina de computar, però aquesta computació encara no està acabada i, en part, haver d'especificar que es tracta d'una màquina i no d'una persona (cal recordar que el terme "computer" encara en aquells moments s'assigna a una persona que fa comptes) denota que en l'imaginari col·lectiu o, si més no en el de McCulloch, encara no és allò evident (allò evident és el que no necessita dir-se i, en el cas de la metàfora computacional com allò evident no començarà a veure's fins la dècada dels 70).

Precisament, en el text McCulloch defensa que és millor entendre el cervell com un procés d'informació, és a dir, que cal deixar enrere la part mecànica i substituir-la per una part digital (i encara que això impliqui més electrònica, el que passarà és que el vocabulari de l'electrònica s'anirà substituint pel vocabulari de la teoria de la informació de Shannon per acabar convertint-se en el vocabulari pròpiament de la informàtica):

> Instead of power, let us think in terms of information conveyed by signals. These signals can be divided into two kinds. First of these is the analogical signal in which the quantity of some variable changes continuously with that which it conveys, like distance in a slide rule or current in a telephonic repeater […]. The second type comprises the logical, or digital, signals which are divided into a few possible quantities whose number in separate places or times is the message to be conveyed—like pins in a cribbage board or dots and dashes in a telegraphic relay.[466]

McCulloch defensarà que el cervell és d'aquest segon tipus, digital, però implementat encara amb relés: «The brain is a logical machine. Each of some ten billion relays has only two states: pulse or no pulse. Each relay is a living cell, shaped something like a vegetable with leaves like a carrot, body like a turnip, and a long thin tap root like alfalfa»[467]. Aquesta perícia descriptiva, que acompanya diferents passatges del text, és el que porta a parlar de la segona hipòtesi de per què aquest article va ser bastant ignorat, com es veurà més endavant.

La barreja entre una visió digital des d'una perspectiva electrònica es va fent evident també en altres moments: «The membrane surrounding this cell is a leaky capacitor, its voltage supported by

465 MCCULLOCH, Warren S. (1949). "The brain as a computing machine" en *Electrical Engineering*, Volume 68, Issue 6, juny 1949, pàg. 492. Consultat el 2 d'agost de 2024 a: https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=6444817

466 *Ídem.*

467 *Ídem.*

the local battery. The rate at which the pulse travels is determined by the distributed resistance, distributed capacity, and distributed source of voltage of the cell, so that cells can be thought of as distributed repeaters»[468]. *Condensadors electrònics* (capacitor), *voltatge*, *bateries*, *resistències* i *repetidors* són les paraules utilitzades per defensar una proposta digital que, mica en mica, farà obsolet aquest mateix vocabulari.

En la comparació entre els punts forts i dèbils tant de la neurona com de la màquina de computar, aquests estan repartits. Per una banda, les neurones són més petites, més barates i requereixen menys energia que qualsevol màquina de computar que fes una mínima part del mateix: «A large building could not house a vacuum tube computer with as many relays as a man has in his head, and it would take Niagara Falls to supply the power and Niagara River to cool it»[469]. Resulta interessant aquesta perspectiva energètica que després recuperarà Kate Crawford en *An Atlas of AI* (2021), obra en la qual analitza tota l'assistència natural, tant de recursos hidràulics, energètics i minerals, com humans i polítics que necessita el projecte IA. Tanmateix, el 1949 la crisi ecològica encara no és el tema del moment i del que es tracta és d'extreure de la natura tot allò que es pugui, també dels seus millors exemplars: «If it cost a million dollars to beget a man, a nerve cell would not cost a mil, and until cathode, grid, and plate can be printed on plastic with only monomolecular films between them, engineers cannot hope to compete with nature».[470]

Per tant, si els enginyers no poden esperar competir amb la naturalesa és perquè, d'alguna forma, l'objectiu és fer-ho, cosa que implica emmirallar-s'hi, és a dir, el sentit o fletxa comparativa va de l'ordinador cap el cervell, no del cervell cap l'ordinador, fins i tot quan el que es posa de manifest són els avantatges que suposa la màquina de computar, entre ells, el nombre estimat d'errors que produeixen ambdós òrgans respectivament: «Brains may seem to be all right when several parts have ceased to function. Any machine like Eniac, which does many things in parallel, is hard to trouble-shoot for the same reason. Parts may fail and answers continue to be computed. But human brains are incomparably worse in this regard»[471]. En això són pitjors, ara bé, també és cert que això només ho determinaria un metge al qual rarament visitaríem, afirma irònicament: «Many a nerve cell can, and does, die and no one knows it till he sees that brain under the microscope. These scattered losses rarely bring us to the doctor».[472]

468 *Ídem.*

469 *Ibídem*, pàg. 493.

470 *Ídem.*

471 *Ibídem*, pàg. 496.

472 *Ídem.*

Aquestes ironies, certes descripcions poc acadèmiques i un estil que costa de saber fins a quin punt cal prendre-se'l seriosament, són la segona hipòtesi de per què aquest text no va ser més citat. El millor exemple d'això són les línies finals de l'article: en el context de tractar les característiques de la neurosi, es planteja què passaria si aquesta comparativa entre el cervell i l'ordinador (en aquest ordre) arribés també a provocar la neurosi a l'ordinador.

> In the neurotic brain you may find no general chemical reaction gone astray, nor any damaged cells, for when activity ceases, regeneration ceases. The most you might expect to find are some changed thresholds or connections— those little invisible differences which each of us acquires by use—the basis of our characters. The more we build negative feedback into machines, the more surely they will have neuroses. These diseases are demons with ideas and purposes of their own. Physicists have been known to curse them but they cannot be exorcised. If, instead of our variety of psychodynamic nonsense, you wish to think sensibly of them I would suggest, in all seriousness, that you start now to prepare a dimensional analysis of gremlins.[473]

Segurament, no cal apel·lar als gremlins ni a l'exorcisme si el que es vol afirmar és que aquesta malaltia és una gran desconeguda (el terme "neurosi" vas ser eliminat el 1994 del *Diagnostic and Statistical Manual of Mental Disorders*, conegut en el camp de la psicologia com el DSM, i que sol marcar la pauta actualitzada per cada versió, si més no en els països occidentals, del diagnòstic i tractament de les malalties mentals que s'hi reconeixen). Tanmateix, McCulloch no se'n pot estar també d'ironitzar sobre el tema.

Conclusió: un dels articles que treballa explícitament per construir la metàfora computacional va ser poc llegit (o poc citat), però citat per un dels pesos pesants, Claude Shannon.

L'actualització del vocabulari: Shannon (1953)

Una de les poques citacions que rebé el text de McCulloch analitzat en l'apartat anterior fou de Claude E. Shannon en un article de 1953 titulat "Computers and Automata" i que va tenir més acollida tot i no ser un document de presentació de resultats d'una nova investigació, sinó quelcom molt més divulgatiu (Crossref: 40; Scorpus: 38; Google Scholar: 179). Es tracta d'un resum de l'estat de la qüestió i una presentació de les noves vies d'investigació en lògica computacional a petició dels editors de la revista *Proceedings of the I.R.E*. Vist retrospectivament, el text planteja problemes de llavors, però fent ús d'un vocabulari que és actual, especialment quan es posa a analitzar diferents projectes d'aprenentatge en màquines (*learning machines*), les estratègies de les quals són processos vigents en aprenentatge automàtic (*machine learning*). Aquests problemes que llavors són línies d'investigació, actualment són algoritmes aplicats en els LLM. Per tant, allò

473 *Ibídem*, pàgs. 496-497.

interessant de l'article no és que digui res nou quant a la metàfora computacional, sinó que el vocabulari que fa servir per dir quelcom molt similar a McCulloch és, llavors, un vocabulari nou.

Després d'una presentació en què defensa, apel·lant a Babbage i a les línies telefòniques, que no s'ha de fer estrany que una màquina de computar tracti problemes que aparentment tenen poc de numèric (*non-numerical computation*), Shannon justifica implícitament la comparació entre el cervell i l'ordinador com una aproximació per tractar el nou concepte de computador programat per un propòsit general (*general-purpose programmed computer*). Les dades que presenta per a la comparació són les de l'article de McCulloch (tot i modificar-li lleugerament la referència a les cascades de Niàgara: el *large building* de McCulloch s'ha convertit en l'Empire State Building, concretament) i la comparació acaba amb un resum similar: entre el cervell i la màquina de computar hi ha diferències de dimensió, en organització estructural, en fiabilitat i en organització lògica, totes a favor del cervell. De fet, la prudència de Shannon es posa de manifest en la següent afirmació, cosa que fa evident que el seu també és encara un projecte de construcció de la metàfora: «Comparisons of this sort should be taken well salted -our understanding of brain functioning is still, in spite of a great deal of important and illuminating research, very primitive. Whether, for example, the neuron itself is the proper level for a functional analysis is still an open question»[474]. I afegeix un comentari, el vocabulari del qual també s'ha recuperat actualment per designar els LLM: «In contrast, our digital computers look like idiot savants»[475]. Aquesta mateixa expressió, *idiot savant*, l'ha fet servir posteriorment Gary Marcus[476], Melanie Mitchell[477], Rodney Brooks[478] o Erik Larson[479] i prové del síndrome de *savant*, que descriu algú amb molta memòria i capacitat de càlcul, però amb una discapacitat per a la resta (l'exemple arquetípic és el personatge de *Rain man* interpretat per Dustin Hoffman).[480]

474 SHANNON, Claude E. (1953). "Computers an Automata" en *Proceedings of the I.R.E.*,Volume: 41, Issue: 10, octubre 1953, pàg. 1235. Consultat el 3 d'agost de 2024 a: https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=4051186

475 *Ídem.*

476 MARCUS, Gary; DAVIS, Ernest (2019). *Rebooting AI. Building Artificial Intelligence We Can Trust*, Nova York, Vintage Books (Penguion Random House LLC), 2020, pàgs. 13, 29, 64, 199.

477 MITCHELL, Melanie (2019). *Artificial Intelligence. A Guide for Thinking Humans*, Londres, Penguin Random House UK, 2020, pàg. 217.

478 FORD, Martin (2018). *"*Rodney Brooks" en *Architects of Intelligence*, Birmingham, Packt Publishing, 2018, pàg. 436.

479 LARSON, Erik J. (2021). *The myth of artificial intelligence: why computers can't think the way we do*, Londres, The Belknap Press of Harvard University Press, 2021, pàgs. 84, 154, 229, 279.

480 TREFFERT, Darold A. (27.05.2009). "The savant syndrome: an extraordinary condition. A synopsis: past, present, future" en *Philos Trans R Soc Lond B Biol Sci*, 364(1522), 27.05.2009, pàgs. 1351-1357. Consultat el 4 d'agost de

Tot i aquesta prudència i a diferència de McCulloch, Shannon fa un passet més cap a la consolidació de la metàfora computacional al presentar una sèrie d'idees i prototips sobre aprenentatge automàtic, no només utilitzant un vocabulari que arrelarà en el camp de la IA, sinó també proposant unes tècniques que són similars a les actuals. Per començar, assumeix que és necessari un procés d'adaptació a l'entorn: «Suppose that an organism or a machine can be placed in, or connected to, a class of environments, and that there is a measure of "success" or "adaptation" to the environment»[481]. És rellevant l'ús de les cometes per marcar que ni la paraula *success* ni la paraula *adaptation* tenen el seu sentit habitual; aquestes cometes són les que han desaparegut en la literatura posterior, fins i tot, per exemple, en els articles de Melanie Mitchell, una de les investigadores més curoses i conscients dels problemes de vocabulari que pateix el projecte IA. Shannon també descriu un segon programa que pretén imitar el comportament reflex dels animals: «The second learning program described by Oettinger is modeled more closely on the conditioned reflex behavior of animals»[482]. S'ha vist que aquest també va ser un projecte que va seguir Rodney Brooks a la dècada dels 80 mentre treballava al MIT. Un tercer projecte descrit per Shannon pretén basar el comportament en l'estudi de patrons anteriors, talment els LLM actuals: «The machine is so constructed as to analyze certain patterns in the players' sequence of choices, and attempt to capitalize on these patterns when it finds them»[483]. Un quart projecte es basava en la integració dels projectes anteriors a través d'una nova màquina: «A third small machine was constructed to act as umpire and pass the information back and forth between the machines concerning their readiness to make a move and the choices made»[484]. La descripció recorda el que es fa en sistemes multi-agents o, simplement, quan s'integren diferents tècniques d'aprenentatge automàtic.

Per tant, el paper de Shannon en la construcció explícita de la metàfora computacional és el de la introducció d'un vocabulari que farà fortuna i que serà el que s'utilitzarà posteriorment, però mentre que Shannon el fa servir des de la prudència (fent servir les cometes, especificant les diferències en la comparació o explicitant que es tracta d'un projecte d'investigació), dues dècades més tard serà l'axioma inicial, la idea que no cal demostrar per evident: el cervell és un ordinador.

2024 a: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2677584/

481 SHANNON, Claude E. (1953). "Computers an Automata" en *Proceedings of the I.R.E.*,Volume: 41, Issue: 10, octubre 1953, pàg. 1238. Consultat el 3 d'agost de 2024 a: https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=4051186

482 *Ibídem*, pàg. 1239.

483 *Ídem.*

484 *Ídem.*

El gran responsable d'aquesta consolidació de la metàfora computacional és von Neumann, qui en fa una anàlisi sistemàtica i qui, pel seu nivell d'influència (formava part primer de RAND Corporation, un organisme independent de defensa dels Estats Units, i després fou un dels primers membres d'ARPA, Advanced Research Projects Agency Network[485]), era especialment escoltat.

La sistematització de la comparació: von Neumann (1957)

El quart intent de construir una comparació explícita entre un ordinador i el cervell (en aquest ordre) el fa John von Neumann el 1957. Es tracta d'un llibre que va aparèixer pòstumament perquè hi va estar treballant des de l'hospital fins poc abans de morir: «The last academic assignment that von Neumann accepted was to deliver and prepare for publication the Silliman lectures at Yale. He worked on that job in the hospital where he died, but he couldn't finish it»[486]. Aquesta conferència, que segons la seva dona li feia molta il·lusió, no la va poder acabar impartint, però un any més tard, va sortir publicada en format de llibre. L'anàlisi d'aquest document servirà aquí per veure els biaixos originals que hi ha en el que més endavant s'anomenarà la metàfora computacional.

Una de les primeres frases del text ja fa evident el grau de provisionalitat del projecte:«Since I am neither a neurologist nor a psychiatrist, but a mathematician, the work that follows requires some explanation and justification. It is an approach toward the understanding of the nervous system from the mathematician's point of view»[487]. Per tant, tampoc aquí l'analogia encara és una analogia de la qual es parteixi, sinó que és la construcció d'un projecte d'interpretació matemàtic del sistema nerviós que acabarà posant les bases per poder fer aquesta analogia per defecte. Aquest caràcter embrionari, von Neumann també el deixa molt clar quan titlla de sobrevaloració el resultat d'aquesta aproximació a l'enteniment. Per altra banda, també especifica què entén i per què amb l'expressió "des del punt de vista matemàtic": «Furthermore, logics and statistics should be primarily, although not exclusively, viewed as the basic tools of ''information theory''».[488]

485 TEMPLE-RASTON, Dina (9.10.2015). "The secretive government agency where 'anything imagined can be tried'" en *The Washington Post*. Consultat el 4 d'agost de 2024 a: https://www.washingtonpost.com/opinions/the-secretive-government-agency-where-anything-imagined-can-be-tried/2015/10/08/3227bc0c-50ce-11e5-933e-7d06c647a395_story.html

486 HALMOS, Paul R. (1973). "The Legend of John von Neumann" en *The American Mathematical Monthly*, Vol. 80, No. 4 (Apr., 1973), pàg. 393. Consultat el 22 de juliol de 2024 a: https://doi.org/10.2307/2319080

487 NEUMANN, John von (1953). *The Computer & the Brain,* New Haven i Londres, Yale University Press, 2012, pàg. 1. Consultat el 26 de juliol de 2024 a: https://ia600707.us.archive.org/3/items/0300181116TheComputerBrain_201901/0300181116_The%20Computer%20Brain.pdf

488 *Ibídem*, pàg. 2.

Ara bé, segurament l'afirmació més interessant d'aquesta introducció prèvia a l'anàlisi és la següent: «I suspect that a deeper mathematical study of the nervous system—"mathematical" in the sense outlined above—will affect our understanding of the aspects of mathematics itself that are involved. In fact, it may alter the way in which we look on mathematics and logics proper»[489]. Per tant, l'objectiu no és que la matemàtica condicioni les facultats del cervell, sinó que és l'objecte d'estudi el que condicionarà l'eina, cosa molt raonable si tenim en compte que, en la majoria d'àmbits pràctics, això és una obvietat: és la tasca a fer la que condiciona l'eina que cal utilitzar, així es tria el martell per clavar un clau, però un tornavís per collar un vis. Aquí von Neumann reconeix que, experimentalment, provarà d'aplicar la matemàtica a l'estudi de l'enteniment. I tot i ser molt raonable, es fa difícil de comprendre en quin sentit pot canviar la manera d'entendre la matemàtica i la lògica, a no ser que sigui precisament per convertir-les en el centre, en fixar-les com a eines preponderants per estudiar l'enteniment, cosa que històricament és el que acabarà passant durant les dècada dels 60 i 70, en què s'assenta el funcionalisme de Putnam i computacionalisme de Fodor i, els anys 80, amb la reintroducció de lingüística generativa de Chomsky (totes aquestes propostes tenen un comú denominador: la teoria de la informació). En curt, mentre que von Neumann sembla defensar que la investigació acabarà portant a un pensar de la lògica (i, estrictament parlant, a ell el porta a això), el que acaba succeint històricament és una lògica del pensar.

Després d'aquesta introducció, von Neumann divideix l'obra en dues parts: la primera part tracta de l'ordinador i fa una explicació del seu funcionament (cal tenir en compte com deuria sonar aquesta explicació el 1957, escassos deu anys més tard de la creació del primer ordinador, l'ENIAC, i per què von Neumann detalla el funcionament del control lògic, el control de la memòria i el seu accés com grans novetats, quan la seva explicació avui en dia es fa a un curs d'introducció a l'estructura de computadors); la segona part tracta del cervell. En vistes a la confecció de la metàfora computacional que s'està forjant, cal observar com l'esforç principal de von Neumann és fer entendre com una màquina digital electrònica, només amb zeros i uns, pot computar i fer-ho de forma més eficient gràcies a la velocitat de l'electricitat, que no pas una màquina analògica mecànica. A mode d'exemple per entendre quina és la principal preocupació de von Neumann, se citen les següents línies:

> To sum up: all these operations now differ radically from the physical processes used in analog machines. They all are patterns of alternative actions, organized in highly repetitive sequences,

489 *Ídem.*

and governed by strict and logical rules. Especially in the cases of multiplication and division these rules have a quite complex logical character.[490]

L'únic element que fa pensar que von Neumann no només està plantejant la comparació sinó que d'alguna manera ja l'està aplicant és l'ús del terme "òrgan" per referir-se a parts del processador (74 vegades utilitza el terme, 52 de les quals a la primera part del text, és a dir, a la part de l'ordinador, quan semblaria poc necessari), especialment a la memòria (*memory organs*), nomenclatura que, tanmateix, no és vigent actualment. Tanmateix, seria una crítica injusta si es té en compte que el concepte llatí d'òrgan (*organum*) precisament és el d'útil o instrument (de fet, en anglès, el segle XII es comença a fer servir per referir-se a l'instrument musical, i no és fins el 1540 en què es veu un ús en el sentit d'una cosa que fa una funció)[491]. Per tant, per prudència no s'afirmarà aquí que von Neumann està fent un ús plenament etològic del terme.

La segona part està dedicada al cervell i des de les primeres línies von Neumann aplica un criteri reduccionista: el cervell és un autòmat el funcionament del qual és *prima facie* digital[492]. Ara bé, al mateix temps que reconeix que en el seu funcionament intervenen aspectes elèctrics, químics i mecànics, assumeix que no cal tenir en compte aquesta diferència en la mesura que, a aquesta escala, el resultat és el mateix: «To sum up: on the usual (macroscopic) scale, electrical, chemical, and mechanical processes represent alternatives between which sharp distinctions can be maintained. However, on the near-molecule level of the nerve membrane, all these aspects tend to merge»[493]. Per poder traçar el paral·lelisme entre l'ordinador i el cervell és imprescindible que els elements químics i mecànics desapareguin de la comparació, ja que l'ordinador que descriu von Neumann és electrònic i funciona amb tubs termoiònics (*vacuum tube* o vàlvula de buit) o transistors. A continuació, per poder mostrar com el cervell també té un comportament digital, redueix el comportament de la neurona al fet que hi hagi o no un impuls: si hi ha impuls, 1; si no hi ha impuls, 0. Aquest aspecte és dels primers que es va criticar i, de fet, la proposta connexionista que defensen Paul i Patrícia Churchland recull la idea dels pesos sinàptics i com emular-los a través de capes ocultes de transistors (anomenades xarxes neuronals i la base dels actuals sistemes d'IA

490 *Ibídem*, pàg. 10.

491 organ (n.). *Online Etymology dictionary.* Consultat el 26 de juliol de 2024 a: https://www.etymonline.com/word/organ

492 NEUMANN, John von (1953). *The Computer & the Brain,* New Haven i Londres, Yale University Press, 2012, pàg. 40. Consultat el 27 de juliol de 2024 a: https://ia600707.us.archive.org/3/items/0300181116TheComputerBrain_201901/0300181116_The%20Computer%20Brain.pdf

493 *Ibídem*, pàg. 42.

generativa), crítica de la qual encara el 1990 s'hauran de defensar. Així ho justifiquen ells en el context del projecte de Carver A. Mead, del California Institute of Technology, qui el 1989 havia creat un retina artificial de silici que simulava la retina d'un gat: «Whether Mead's program could be sustained to build an entire artificial brain remains to be seen, but there is no evidence now that the absence of biochemicals renders it quixotic»[494]. En efecte, von Neumann n'és plenament conscient i accepta que aquest *prima facie* acaba sent una mica més complex: «Let me add a few words regarding the qualifying ''prima facie.'' The above description contains some idealizations and simplifications, which will be discussed subsequently. Once these are taken into account, the digital character no longer stands out quite so clearly and unequivocally»[495]. Aquesta prudència és la que, vint anys més tard, quan la metàfora computacional ja està completament integrada en la vida quotidiana, desapareixerà.

La resta de la segona part és un *tour de force* per aconseguir trobar similituds entre el cervell i l'ordinador, especificant en la comparació les diferències d'ordre de magnitud en capacitat de cada un d'ells. Així, a nivell de velocitat, posa en valor que les connexions de l'ordinador són molt més ràpides que les del cervell, mentre que quant a dimensions i capacitat de memòria, les neurones són molt més petites que el xips del moment i poden emmagatzemar més informació. Tant un com altre valor (la velocitat, la dimensió o la memòria), von Neumann els explica com a valors positius: així, el fet que les connexions elèctriques vagin més ràpides que les neuronals, és una aspecte positiu dels ordinadors, mentre que el fet que no siguin més petits que una neuroa, és un aspecte positiu de la naturalesa. Aquest prejudici és al mateix temps l'estímul que marcarà com ha de seguir aquest projecte si es vol aconseguir realment una equiparació completa entre ordinador i cervell, i no és altra que el que descriu Gordon Earle Moore en el que s'ha anomenat la llei de Moore (que més que una llei, és un objectiu de treball): cal fer transistors més petits per poder fer circuits amb més capacitat i més emmagatzematge. Actualment, quan s'està arribant al límit de la possibilitat de miniaturització dels microxips a causa de no poder sobrepassar l'àtom de silici, s'han estès dues idees: per una banda, pot ser el final del projecte IA; per altra, cal trobar mecanismes en arquitectura de computadors que permetin superar aquest límit. Una de les idees que treballa aquesta segona proposta se l'ha anomenat *More Moore* i es basa en la construcció en tres

494 CHURCHLAND, Patricia S.; CHURCHLAND, Paul M. (1990). "Could a Machine Think?" en *Scientific American*, 262, 1, 1990, pàg. 37. Consultat el 22 d'agost de 2024 a: http://www.jstor.org/stable/24996642.

495 NEUMANN, John von (1953). *The Computer & the Brain,* New Haven i Londres, Yale University Press, 2012, pàg. 44. Consultat el 27 de juliol de 2024 a: https://ia600707.us.archive.org/3/items/0300181116TheComputerBrain_201901/0300181116_The%20Computer%20Brain.pdf

dimensions de la circuiteria habitualment plana; una altra és fer servir algun material que permeti crear portes lògiques més petites, com el grafè; i finalment, hi ha la promesa de la computació quàntica[496].

Les últimes idees que von Neumann escriu en aquest text entren a considerar si el tractament quantitatiu (aritmètic i lògic) del cervell té o no té sentit: «Thus one would expect that the arithmetical part of the nervous system exists and, when viewed as a computing machine, must operate with considerable precision»[497]. I en fer-ho, fa un pas en fals, que podria considerar-se una inversió de la metàfora computacional: «As pointed out before, we know a certain amount about how the nervous system transmits numerical data»[498]. La fal·làcia d'aquesta afirmació té a veure amb certa concepció realista de la matemàtica, ja que, estrictament parlant, no és que el sistema nerviós transmeti dades numèriques, sinó que es quantifiquen numèricament una sèrie de reaccions del sistema nerviós amb l'objectiu de construir un model. Per tant, l'afirmació inicial de von Neumann sobre com podria alterar aquest estudi la manera de veure l'essència de la matemàtica i la lògica, o manca de sentit o vol dir això: els objectes matemàtics són reals, doncs operen també en el cervell[499].

Aquesta és, de fet, l'última idea que von Neumann tracta en les línies finals del text: «Just as languages like Greek or Sanskrit are historical facts and not absolute logical necessities, it is only reasonable to assume that logics and mathematics are similarly historical, accidental forms of expression»[500]. Tanmateix, això no vol dir que el sistema nerviós no sigui matemàtic o lògic, sinó que no necessàriament ho són respecte a la matemàtica i la lògica que s'està fent servir actualment: «Thus the outward forms of *our* mathematics are not absolutely relevant from the point of view of evaluating what the mathematical or logical language *truly* used by the central nervous system

496 KNOT MAY, Rosalie (25.05.2023). "More than Moore: the next steps for the semiconductor industry" en *Delmic*. Consultat el 27 de juliol de 2024 a: https://blog.delmic.com/more-than-moore-the-next-steps-for-the-semiconductor-industry

497 NEUMANN, John von (1953). *The Computer & the Brain,* New Haven i Londres, Yale University Press, 2012, pàg. 77. Consultat el 27 de juliol de 2024 a: https://ia600707.us.archive.org/3/items/0300181116TheComputerBrain_201901/0300181116_The%20Computer%20Brain.pdf

498 *Ídem.*

499 MADDY, Penelope (1993). *Realism in mathematics (A review)*, Londres, Oxford University Press, 1993. Consultat el 27 de juliol de 2024 a: https://web.archive.org/web/20180726044843id_/http://www.ams.org/journals/bull/1995-32-01/S0273-0979-1995-00552-5/S0273-0979-1995-00552-5.pdf

500 *Ibídem*, pàg. 82.

is»[501]. És a dir, l'estudi del cervell en un sentit matemàtic ha posat de manifest, segons von Neumann, que el cervell (com qualsevol element del món) és matemàtic. Per tant, l'estudi de la matemàtica, és a dir, de l'ordre i connexió de les idees, permetrà conèixer l'ordre i connexió del món, tesi que no només recupera la idea de Spinoza d'*ordo et connexio*, sinó que és una de les bases conceptuals de tot el projecte modern, des de Descartes a la IA forta, i un dels seus defensors, Ned Block.

5.2.2 L'estabilització de la metàfora

A continuació s'estudien dos textos que denoten que la metàfora computacional ja és operativa amb normalitat, en un primer moment només en el camp de la recerca informàtica i, posteriorment, també en altres camps. El primer text és de Joseph Weizenbaum (1923-2008) i l'escriu el 1972; el segon text és de Richard Boyd (1942-2021) i l'escriu el 1979. El primer avisa d'un perill futur: l'ús de la metàfora computacional és enganyós i caldria fer pedagogia per evitar que proliferi el seu mal ús. El segon assumeix que la metàfora computacional és la norma en camps com la psicologia. En ambdós casos, la metàfora no és el que s'ha de justificar, sinó allò a denunciar o allò que es fa servir amb certa naturalitat. En ambdós casos, és evident que és sempre present.

Weizenbaum: *side effects of technology* (1972)

La primera referència que en aquest treball s'ha pogut trobar dels perills de l'ús de la metàfora computacional és la de Joseph Weizenbaum en un article publicat el maig de 1972 a la revista *Science* amb el títol "On the Impact of the Computer on Society". Weizenbaum era, just fins abans de la publicació d'aquest article, una veu respectada especialment entre la comunitat enginyeril del MIT, on treballava des de 1963: venia de treballar a la General Electric Co. i havia publicat tres articles, "The GE-100 Data Processing System" (1958), "Knotted List Structures" (1961) i "Symmetric List Processor" (1963), que li havien fet guanyar certa fama com a programador competent capaç de crear un nou llenguatge de programació amb una estructura per nusos (*knotted list structures*) que acabaria sent la base computacional d'ELIZA, el primer programa d'ordinador amb el qual un usuari final es podia comunicar en anglès. Ja treballant al MIT, va publicar "ELIZA —a computer program for the study of natural language communication between man and machine" (1966), que el va catapultar a la fama, i també a la necessitat d'agafar-se dos anys sabàtics entre 1973 i 1975. Durant aquests anys sabàtics, Weizenbaum coneix Mumford, Chomsky i Steven Marcus (un crític literari), als quals agraeix la lectura del seu manuscrit, i es forma filosòficament amb Putnam i Dennett («an outstanding young philosopher from Tufts University»[502]), als quals agraeix la seva paciència amb la seva (la de Weizenbaum) ingenuïtat filosòfica. Anys més tard, quan

501 *Ibídem*, pàg. 83. La cursiva és de von Neumann.

rememora el seu pas de l'empresa a la universitat, així com la necessitat de crear un programa com ELIZA, ho lamenta:

> P.- Usted pasa de la industria y la empresa privada a la universidad, ¿cómo valora este cambio?
> R.- Cuando era estudiante en la universidad me interesaban los temas sociales y políticos. Cuando tuve la libertad que te da un contrato académico, que no es ilimitada pero sí considerable, intenté pensar en temas y proyectos que apuntasen en esa dirección. Pensé que un problema fundamental a abordar era que los ordenadores entendiesen el lenguaje natural y nos pudiéramos comunicar con ellos en inglés, alemán, español, etc. Ahora pienso que fue un error. El error no fue trabajar en ello, sino pensar que era un requisito necesario para que los ordenadores trabajasen en temas sociales.[503]

Weizenbaum explica que l'error el va començar a intuir quan la seva secretària, qui l'havia vist passar-se hores programant ELIZA, li va demanar si podia marxar de la sala doncs estava tenint una conversa privada:

> My secretary watched me work on this program over a long period of time. One day she asked to be permitted talk with the system. Of course, she knew she was talking to a machine. Yet, after I watched her type in a few sentences she turned to me and said "Would you mind leaving the room, please?" I believe this anecdote testifies to the success with which the program maintains the illusion of understanding. However, it does so, as I've already said, at the price of concealing its own misunderstandings.[504]

A partir d'aquest moment, Weizenbaum capgira la seva carrera acadèmica i passa a convertir-se en l'abanderat de la lluita contra la colonització digital i la necessària col·laboració científica per dur a terme el projecte, primer des d'un vessant pedagògic, com en la seva obra principal, *Computer Power and Human Reason* (1976), però mica en mica de forma més activista, com en "Not Without Us" (1987), també en la seva última entrevista publicada, *Islands in the Cyberstream. Seeking Havens of Reason in a Programmed Society* (2006) o en el documental pòstum *Plug & Pray* (2010).

---

502 WEIZENBAUM, Joseph (1976). *Computer Power and Human Reason*, Londres, Penguin Books Ltd, New Ed, 1984, pàg. xi.

503 WEIZENBAUM, Josep (1993). "Entrevista a Joseph Weizenbaum" en *Telos*, núm.38, Fundación Telefónica. Consultat el 30 de juliol de 2024 a: https://telos.fundaciontelefonica.com/archivo/numero038/entrevista-a-joseph-weizenbaum/

504 WEIZENBAUM, Joseph (1967). "Contextual Understanding by Computers" en *Communications of the ACM*, Volume 10, Number 8, August 1967, pàgs. 474-478. Consultat el juliol de 2019 a: https://doi.org/10.1145/363534.363545

Ara bé, la primera mostra d'aquest canvi apareix, com s'ha dit al principi de l'apartat, en l'article "On the Impact of the Computer on Society" (1972). En aquell moment, per a Weizenbaum l'impacte social de l'ordinador encara no és tan preocupant com transmeten els mitjans de comunicació (el *hype* del moment era aquest), però a ell li interessa l'impacte a llarg termini que creu que tindrà: «[...] the direct societal effects of any pervasive new technology are as nothing compared to its much more subtle and ultimately much more important side effects. In that sense, the societal impact of the computer has not yet been felt»[505] –aquesta mateixa idea és la que, uns 30 anys més tard, va fer famós a Roy Amara amb la llei d'Amara, com s'ha vist al capítol 4. Tanmateix, a Weizenbaum aquesta idea no li serveix com a conclusió, sinó com a idea inicial per començar a construir una teoria sobre el paper de la ciència i tecnologia en la societat i, en concret, per analitzar si l'invent de l'ordinador pot tenir la mateixa rellevància que se sol assignar al microscopi en el canvi al paradigma modern. Comença per separar la teoria de la pràctica (*Theory versus Performance*) i assumeix que un model teòric no es pot sustentar només pels seus resultats pràctics (funciona o no funciona, prediu encertadament o no), sinó també per la seva capacitat explicativa del fenomen. Aquesta diferència, que en altres camps és òbvia, Weizenbaum creu que queda amagada sota l'aparença objectiva que aporta, per sistema, la computació:

> Perhaps by the end of the present decade, computer systems will exist with which specialists, such as physicians and chemists and mathematicians, will converse in natural language. And surely some part of such achievements will have been based on other successes in, for example, computer simulation of cognitive processes. It is understandable that any success in this area, even if won empirically and without accompanying enrichments of theory, can easily lead to certain delusions being planted. Is it, after all, not terribly tempting to believe that a computer that understands natural language at all, however narrow the context, has captured something of the essence of man? Descartes himself might have believed it. Indeed, by way of this very understandable seduction, the computer comes to be a source of philosophical dogma.[506]

Més enllà de l'optimisme que impregna la primera línia –fins el 2022, amb la creació dels GPTs entrenats amb *big data*, seria difícil defensar això i, fins i tot actualment, els problemes de les al·lucinacions i la manca d'escalabilitat permeten posar-ho en dubte com ho fan els autors escèptics estudiats al capítol 4–, la clau està en com explica l'aparició d'aquest nou dogma: en el cas de la computació, la *performance* no pot ser criteri per acceptar una teoria, ja que entre el bon

505 WEIZENBAUM, Joseph (1972). "On the Impact of the Computer on Society. How does one insult a machine?" en *Science*, vol. 175, maig 1972, pàg. 609. Consultat el 22 d'agost de 2024 a: https://doi.org/10.4000/socio-anthropologie.13631

506 *Ibídem*, pàg. 611.

funcionament d'un ordinador i una teoria lingüística (o una teoria de la ment) no hi ha cap relació causal (a diferència del disseny d'avions i l'aerodinàmica), sinó un vincle metafòric, la metàfora computacional –tot i que en aquest article diverses vegades l'anomena "metàfora tecnològica", la vincula directament amb certa idea de la computació i llenguatges de programació, per tant, a tots els efectes està tractant la metàfora computacional. Els motius pels quals la *performance* d'un programa d'ordinador no pot servir per mesurar el grau d'encert d'un model teòric són diversos:

1. Un codi d'un programa no tradueix habitualment ni necessàriament un model teòric, sinó que consisteix en un conjunt de tècniques que donen resultat (el terme que sol utilitzar Weizenbaum és *patchwork*)[507]. Si aquesta frase ja era vàlida el 1972, quan les línies de codi d'un programa eren de pocs milers, actualment, amb les anomenades *black boxes* dels GPTs, encara és més evident.

2. Un programa d'ordinador que simuli el comportament d'un òrgan humà no implica que el funcionament del simulador sigui pels mateixos motius que el funcionament de l'òrgan (malgrat ambdós processos es puguin traduir a àlgebra booleana).[508]

3. L'organització de components computacionals que simuli l'organització d'òrgans humans no permet explicar ni simular com el canvi de l'organització computacional afectaria al mateix canvi en l'organització orgànica. I això, encara menys, és extrapolable als éssers humans en conjunt.[509]

A Weizenbaum no li preocupa tant que això estigui passant en el món acadèmic, sinó que això acabi traspassant a la societat civil (tot i que ell mateix considera que encara no ha succeït: «The computing metaphor is as yet available to only an extremely small set of people»[510]). En cas que passi, o quan passi, preveu que hi hagi efectes col·laterals negatius a llarg termini pels següents motius:

1. Un programa d'ordinador no solucionarà problemes socials propis dels humans (un exemple que farà servir més endavant és el de la lectura: no és comprant un ordinador per cada alumne que es resoldrà el problema de comprensió lectora, sinó analitzant com és que hi ha famílies que no llegeixen amb els seus infants a casa). Per tant, si es tendeix a donar respostes d'aquest tipus, els problemes s'ocultaran darrere de capes de tecnologia.

507 *Ibídem*, pàg. 612.

508 *Ídem.*

509 *Ídem.*

510 *Ibídem*, pàg. 613.

2. Hi ha un perill que es delegui en els ordinadors, no només la solució dels problemes en tant que *problem-solvers*, sinó també la formulació de les preguntes: «What is wrong, I think, is that we have permitted technological metaphors, what Mumford calls the "Myth of the Machine," and technique itself to so thoroughly pervade our thought processes that we have finally abdicated to technology the very duty to formulate questions».[511]

3. Utilitzar ordinadors per encarar certs problemes és una forma de delegar la responsabilitat: «No human is any longer responsible for "what the machine says"»[512]. Aquesta característica l'assenyalarà novament Cathy O'Neil quan descriu com, no només els bancs amb les hipoteques, sinó també els jutges amb les sentències, els costa contradir la proposta que fa un algoritme amb IA: «If we back away from them and treat mathematical models as a neutral and inevitable force, like the weather or the tides, we abdicate our responsibility»[513]. En l'apartat dedicat a la relació entre la metàfora computacional i l'etologia s'intentarà explicar quines són les bases per a aquesta delegació de responsabilitat.

Però Weizenbaum també assumeix que poden haver-hi efectes positius (aquesta actitud equidistant ja no apareixerà en la següent obra i, molt menys, a partir de 1980), entre ells, el fet que l'ús d'una nova metàfora pot enriquir el marc intel·lectual i obrir la porta a noves preguntes, és a dir, ofereixi una nova perspectiva del món. Ara bé, com amb qualsevol metàfora, Weizenbaum pensa que cal evitar caure en la desídia de prendre-se-la al peu de lletra i com a solució per a tot: «Indeed, the very effectiveness of a new metaphor may seduce lazy minds to adopt it as a basis for universal explanations and as a source of panaceas»[514]. D'alguna manera, com Wiener o Shanahan, l'actitud adequada davant de la metàfora computacional és entre la vigilància constant i no prendre-se-la massa seriosament.

Boyd i els usos de la metàfora computacional en el camp de la psicologia (1979)

Com s'ha vist en el primer apartat d'aquest capítol, Richard Boyd dialoga amb Max Black sobre l'ús de les metàfores en la ciència. Boyd defensa que, en alguns casos, les metàfores tenen

511 *Ibídem*, pàg. 611.

512 *Ibídem*, pàg. 613.

513 O'NEIL, Cathy (2016). *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy*, New York (USA), Crown Publishing Group, 2016, pàg. 184. Consultat el 2 d'agost de 2024 a: https://edisciplinas.usp.br/pluginfile.php/7574239/mod_resource/content/1/%28FFLCH%29%20LIVRO%20Weapons%20of%20Math%20Destruction%20-%20Cathy%20ONeal.pdf

514 WEIZENBAUM, Joseph (1972). "On the Impact of the Computer on Society. How does one insult a machine?" en *Science*, vol. 175, maig 1972, pàg. 613.

una funció constitutiva, en concret, quan el vocabulari tècnic d'un camp encara no està suficientment assentat i requereix d'unes estructures comunicatives més obertes (*open-ended*) que es puguin assentar en el món per juntures encara no molt clares. Per justificar aquesta afirmació, Boyd posa l'exemple de l'ús de la metàfora computacional en el camp de la psicologia, i enumera tots els casos que ell ha detectat que es fa servir:

> 1. the claim that thought is a kind of "information processing," and that the brain is a sort of "computer";
> 2. the suggestion that certain motoric or cognitive processes are "preprogrammed";
> 3. disputes over the issue of the existence of an internal "brain-language" in which "computations" are carried out;
> 4. the suggestion that certain information is "encoded" or "indexed" in "memory store" by "labeling," whereas other information is "stored" in "images";
> 5. disputes about the extent to which developmental "stages" are produced by the maturation of new "preprogrammed" "subroutines," as opposed to the acquisition of learned "heuristic routines," or the development of greater "memory storage capacities" or better "information retrieval procedures";
> 6. the view that learning is an adaptive response of a "self-organizing machine";
> 7. the view that consciousness is a "feedback" phenomenon.[515]

Boyd defensa que, encara que aquests usos no siguin de vital importància per a la teoria psicològica, han provocat la creació de nou vocabulari: «These metaphors have provided much of the basic theoretical vocabulary of contemporary psychology (Neisser, 1966; G. A. Miller, 1974)»[516]. I per acabar de fonamentar la seva afirmació, menciona el funcionalisme i una llarga llista d'autors i obres: Block & Fodor, 1972; Block 1977; Boyd, 1980; Fodor, 1965, 1968; Lewis, 1971; Putnam, 1967, 1975b, 1975f; Shoemaker, 1975b. L'objectiu de Boyd no és afirmar que totes elles combreguen exactament amb la mateixa idea, sinó precisament provar la seva teoria sobre l'estat d'obertura d'aquest tipus de metàfores, que serveixen tant per qui defensa que la ment és el *software* del cervell, com qui defensa una visió més dèbil dins el projecte IA.

En qualsevol cas, per l'argument que aquí s'està construint, l'exemple de Boyd serveix per constatar que en vint anys, la metàfora computacional ja no és un projecte, sinó la base de la qual es parteix, una expressió consolidada, fins i tot, un camp d'interès per ell mateix, com així també mostren els números bruts de referències de Google Acadèmic per dècades:

---

515 BOYD, Richard N. (1979). "*Metaphor and theory change: What is metaphor" a metaphor for?" en en *Metaphor and thought,* Editor Ortony A, Cambridge, MA, Cambridge University Press, 1979, pàg. 486.

516 *Ibídem*, pàg. 487.

Dècada	Nombre de referències “computer metaphor”	Δ	Nombre de referències “computer” & “metaphor”	Δ	Nombre de referències de “metaphor”	Δ	Nombre de referències de “computer”	Δ
1930-1940	0		77		8.180		12.800	
1940-1950	1		105	1,36	11.600	1,42	14.100	1,10
1950-1960	2	2,00	282	2,69	15.900	1,37	20.300	1,44
1960-1970	5	2,50	2.532	8,98	16.800	1,06	436.000	21,48
1970-1980	218	43,60	10.300	4,07	23.100	1,38	579.000	1,33
1980-1990	862	3,95	16.700	1,62	112.000	4,85	602.000	1,04
1990-2000	1.640	1,90	111.000	6,65	59.600	0,53	975.000	1,62
2000-2010	2.260	1,38	38.600	0,35	94.500	1,59	1.130.000	1,16
2010-2020	2.840	1,26	48.400	1,25	32.600	0,34	1.220.000	1,08

El resultats de la segona columna és el nombre de vegades que apareix l’expressió “computer metaphor” entre cometes, per tant, com una expressió en si mateixa, mentre que els resultats de la tercera representen la variació d’una dècada respecte a l’anterior (al igual en les columnes 5, 7 i 9). Els resultats de la quarta columna reflecteixen el nombre de vegades que apareixen els dos termes, units o desvinculats, en el mateix text. Els resultats de la sisena i la octava columna, que reflecteixen el nombre de vegades que apareix la paraula “metaphor” i la paraula “computer” respectivament, es poden interpretar com dades de control, i pretenen mostrar com el resultat global aporta certa perspectiva, és a dir, que la desviació és proporcional a totes les columnes.

Tenint en compte això i més enllà d’altres consideracions que es puguin fer, com per exemple que en la dècada de 1930 la majoria de referències que contenen les paraules “metaphor” i “computer” en el mateix text són de l’àmbit de la filologia i no de la computació, i també els possibles errors en la datació de textos, sí que la taula, i més fàcilment el gràfic de sota, permet veure que el *boom* de la computació sorgeix a la dècada dels 60, mentre que el camp d’investigació de la metàfora computacional (“computer metaphor”) esclata a la dècada dels 70, tot i que el tema, o si més no la coincidència de parlar de metàfores i ordinadors en un mateix text, ja comença a notar-se una dècada abans.

En aquest sentit, crida l'atenció un article poc conegut de C.C. Anderson, un autor també poc conegut, titulat "The latest metaphor in Psychology" de 1957, en el qual ja preveu un possible perill de l'aplicació de la metàfora computacional: «In view of so many parallels, the danger clearly lies in a tendency to anthropomorphise the machine and mechanise the man»[517]. Anderson critica que la psicologia no disposi d'una metàfora duradora que li permeti assentar les teories durant períodes llargs de temps, i defensa que una de les metàfores que pot prendre aquest rol és una novíssima idea de McCulloch i de Wiener: «We may ask whether the new metaphor passes muster on these criteria. The brain is a calculating machine»[518]. Un dels criteris que proposa per avaluar si una metàfora pot ser productiva és si la matemàtica que s'aplica en el camp d'origen de la metàfora es pot aplicar en el camp de destí. Posa d'exemple els mètodes matemàtics de Fourier, aplicats en la conducció de la calor, i com després Lord Kelvin els introdueix en l'electroestàtica de forma anàloga. La proposta d'Anderson és verificar si l'analogia entre el cervell i un màquina de calcular (encara no està consolidat si és de calcular o de computar la màquina en qüestió) compleix amb aquesta sèrie de condicions i si val la pena fer-la servir en psicologia: «At once we may say that this metaphor satisfies the first and third positive criteria in that the appropriate mathematical methods have been applied (though not without some criticism, by Cherry among others)»[519]). El 57 encara no és el moment, però a la dècada dels 70 ja està plenament consolidat el seu ús.

517 ANDERSON, C.C. (1957). "The latest metaphor in Psychology" en *Dalhousie Review*, Volume 37, Number 2, 1957, pàg. 182. Consultat el 2 d'agost de 2024 a: https://dalspace.library.dal.ca//handle/10222/58774

518 *Ibídem*, pàg. 180.

519 *Ídem.*

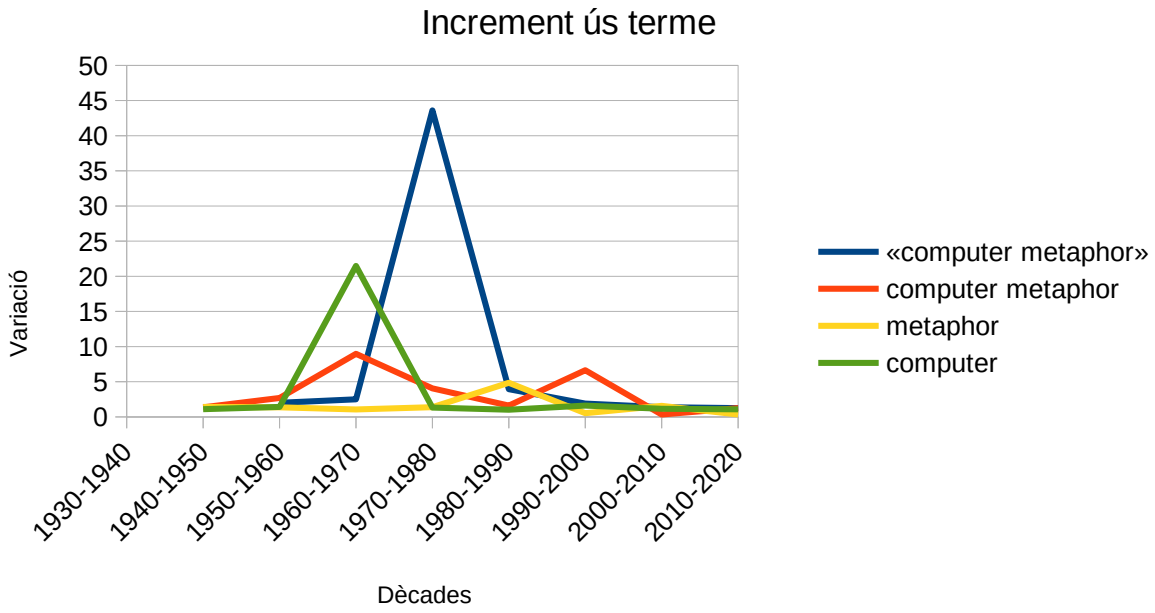

Figura 11: Comparativa ús terme "metàfora computacional"

També és evident que l'interès pels ordinadors és molt superior que l'interès per les metàfores en general i, només a la dècada dels 90, sembla que es posa de moda el combinar els dos termes en un mateix text. També seria interessant investigar què passa a la dècada dels 80 perquè hi hagi un increment significatiu de l'estudi de les metàfores (o s'utilitzi més aquest terme que no pas la dècada anterior).

5.2.3 La deconstrucció de la metàfora computacional

Com tot procés, també la metàfora computacional passa per moments de revisió i/o crisi, i a partir dels anys 80 es comencen a donar dos fenòmens paral·lels que denoten que aquesta estructura, de tan arrelada, cal desocultar-la per revisar els seus implícits, o cal superar-la per assentar una nova metàfora més útil. Això primer, desocultar-la, sembla que fa Daniel Dennett (1942-2024) en un text de 1984; això segon, intentar superar-la, és el que pretén Steven Pinker (1954-) en un text de 1997.

Dennett i el rol de la metàfora computacional en entendre la ment (1984)

Consolidada la metàfora computacional com una línia d'investigació o, com a mínim, una àrea d'interès, Daniel Dennett li dedica un article titulat "The Role of the Computer Metaphor in Understanding the Mind" (1984). El seu és un bon exemple, no només que la metàfora ja està plenament assentada tant en l'àmbit de la computació com en el de la psicologia, sinó que ja es pot

començar a deconstruir, és a dir, a revisar quines peces la formen i quina és la seva funció. I per fer-ho, cal fer aparèixer els problemes filosòfics que amaga. Aquest sembla ser l'objectiu de Dennett i no tant donar algun tipus de resposta o aportar un nou argument: aclarir, explicitar i plantejar el problema.

Dennett comença amb una anècdota sobre l'equivalent dificultat per saber com funciona el cervell només obrint un crani i saber com funciona un ordinador només mirant els xips. Aquesta anècdota li permet descriure les quatre maneres habituals d'encarar aquesta dificultat: el dualisme, que diu que la ment és quelcom diferent i d'un material diferent al del cervell; el misticisme sobre el cervell orgànic, que diu que la ment ha de ser el cervell, però que és essencialment inexplicable; la proposta *bottom-up* / *top down* pròpies de la investigació científica, que pretenen explicar el funcionament del cervell per molt inescrutable que sigui; i l'estratègia de la idealització teorètica (*the strategy of theoretical idealization*). Aquesta és la que analitzarà i, en el fons, utilitzarà en aquest article.

Descriu aquesta estratègia remuntant-se a les propostes epistemològiques clàssiques com la de Descartes, però contraposant-li les propietats inherents a la computació, és a dir, una teoria de la ment ha de poder ser implementable, per tant, ha de complir les següents tres condicions: ha de poder-se explicar mecànicament; ha de suposar una ment finita (en contraposició a les propietats de la *res cogitans* cartesianes, diu Dennett); i ha de donar respostes en un temps raonable (no pot requerir de sis meditacions ni estendre's *sub specie aeternitatis*). Aquestes tres característiques converteixen l'epistemologia clàssica en una epistemologia d'androide (Dennett atribueix el nom a Clark Glymour, qui tanmateix, té la impressió que el projecte IA no és res més que positivisme lògic dut a terme per uns altres mitjans[520]).

Establertes aquestes condicions, Dennett presentarà una doble tesi, aparentment contradictòria (però només aparentment): si per una banda Searle té raó quan diu que aquesta ment només la pot fer un cervell com aquest, és a dir, humà, al mateix temps el problema no és el cervell sinó les funcionalitats que aquest tingui, doncs això és el màxim que es pot predicar de la ment sense caure en un misticisme: «So it may very well turn out that the only way one can achieve the information-handling prowess of a human brain (in real time) is by using-a human brain!»[521]. Per tant, l'estratègia d'analitzar la ment com un procés d'intercanvi de símbols, és a dir, com l'habitació xinesa de Searle (equivalent al *software*), permet tant evitar els problemes del dualisme (en concret,

520 DENNETT, Daniel C.(1984). "The Role of the Computer Metaphor in Understanding the Mind" en *Annals of the New York Academy of Sciences,* Volume 426, Issue 1, pàg. 267. Consultat el 4 d'agost de 2024 a: https://nyaspubs.onlinelibrary.wiley.com/doi/abs/10.1111/j.1749-6632.1984.tb16524.x

521 *Ibídem*, pàg. 269.

el problema del *homunculi*) com integrar les propietats d'una epistemologia d'androide: és mecànica, és finita i és temporal. Llavors, hauria de ser possible construir un cervell de silici que repliqués totes les connexions del cervell humà i, per tant, fos una ment. Tanmateix, sembla que una ment així mancaria d'intencionalitat i consciència (a no ser que emergissin misteriosament, per tant, místicament). Així doncs, no té sentit intentar replicar els circuits del cervell en silici, doncs el cervell és l'únic material en el qual es pot implementar una ment d'aquest tipus: «So it might turn out after all that the only way to have a mind like ours is to have a brain like ours, composed of the same organic materials, organized in roughly the same way»[522]. Però, al mateix temps, una bona manera d'estudiar la ment (i potser l'única manera) és fer-ho per les seves funcionalitats, és a dir, com un procés d'intercanvi de símbols, encara que, fins i tot si s'arribés a poder replicar un cervell de silici, aquest mancaria d'una ment. Allò que podria semblar contradictori és paradoxal, com ho és un oximoron que es pogués resumir així: allò que cal fer, se sap que no permetrà aconseguir el que es vol aconseguir, però és l'únic que es pot fer si es vol aconseguir el que no es pot aconseguir; és a dir, un camí que no porta enlloc, però pel qual cal passar.

Per tant, el funcionalisme i, després, el computacionalisme, serien noves maneres de fer epistemologia des d'un vessant més restringit (si més no, restringit per les tres propietats d'una epistemologia d'androide), més humà (menys diví), amb la possibilitat d'acabar articulant una teoria de la ment que serveixi per ajudar a investigar el cervell: «All this shows, of course, is a certain possibility in principle. It does not establish how or why or in what ways a human mind might be like a computer. All it shows is that in principle we can have a theory of the mind that is mechanistic, finite, operates in real time, and is not a perpetual motion machine».[523]

Així doncs, des del punt de vista de la metàfora computacional, aquesta no només serviria al projecte IA computacionalment, sinó que també podria tenir un sentit dins del projecte neurocientífic, doncs estaria aportant una sèrie d'idees, és a dir, una sèrie de noves teories de la ment restringides per les propietats derivades d'una implementació en silici, que podrien ser d'ajuda per entendre el funcionament del cervell. En altres paraules, la metàfora computacional no seria una estratègia informàtica per fer ordinadors més potents a base d'imitar el cervell, sinó que seria una estratègia neurocientífica per copiar característiques dels ordinadors amb l'objectiu d'explicar el funcionament del cervell.

En aquest treball s'ha anomenat això la inversió de la metàfora computacional i s'exemplifica de la següent manera: si en un inici el cervell era el model i l'ordinador l'objecte que la intentava replicar (l'ordinador era com un cervell), ara el model és l'ordinador i és el cervell el que s'hi ha

522 *Ídem.*

523 *Ibídem*, pàg. 272.

d'assemblar (el cervell és com un ordinador). En el següent capítol s'analitzaran les conseqüències d'aquesta inversió.

Una ressenya d'aquest article feta per Robert L. Stout sembla coincidir amb aquesta interpretació, doncs assenyala la manca d'interès de l'article en vistes a la millora computacional, en canvi, sí que li atorga algun interès des del punt de vista filosòfic :

> Because the author dwells primarily on the implications of AI and cognitive science for philosophy, rather than vice versa, information scientists looking for possible new directions for research will find little guidance in this paper. For those with a philosophical bent, however, the paper provides a strong presentation of how AI and cognitive science are seen from a modern positivist perspective.[524]

Per tant, la metàfora computacional ja no té una funció per millorar un ordinador, sinó per canviar un altre camp, en aquest cas, el de la filosofia, però també veurem com influeix en altres camps, com la psicologia o la pedagogia.

La metàfora computacional i el computacionalisme: Steven Pinker (1997)

Una altra mostra que la metàfora computacional és d'allà on es parteix, és a dir, quelcom evident que no cal demostrar, és l'esforç que fa Steven Pinker (1954) per separar la seva proposta, el computacionalisme, de la metàfora computacional. En una obra ja clàssica, *How the mind works* (1997), Pinker defensa la següent tesi: «The mind is a system of organs of computation, designed by natural selection to solve the kinds of problems our ancestors faced in their foraging way of life, in particular, understanding and outmaneuvering objects, animals, plants, and other people»[525]. Per tant, és un plantejament que utilitza la teoria de l'evolució com a fonament per justificar el funcionament del cervell i, a partir d'aquí, el comportament humà (aquesta idea s'emmarca dins la denominada psicologia evolutiva). Pinker desglossa el conjunt d'idees que hi ha sota el seu plantejament (se separen per punts per ser més clars):

A) The mind is what the brain does; specifically, the brain processes information, and thinking is a kind of computation.

B) The mind is organized into modules or mental organs, each with a specialized design that makes it an expert in one arena of interaction with the world.

C) The modules' basic logic is specified by our genetic program.

524 STOUT, Robert L. (1984). "Review of «The role of the computer metaphor in understanding the mind»" en *ACM Digital Library*. Consultat el 5 d'agost de 2024 a: https://dl.acm.org/doi/10.5555/4959.4979

525 PINKER, Steven (1997). *How the mind works*, Londres, Penguin, 1998, pàg. 21.

D) Their operation was shaped by natural selection to solve the problems of the hunting and gathering life led by our ancestors in most of our evolutionary history.

E) The various problems for our ancestors were subtasks of one big problem for their genes, maximizing the number of copies that made it into the next generation.[526]

Ara bé, cada una d'aquestes afirmacions Pinker les matisarà per tal que el reduccionisme que aparentment defensen no sembli tal reduccionisme. Per fer-ho utilitza una tècnica d'extensió opaca (en el sentit que la reducció d'una reducció és una extensió, però opaca en el sentit de poc clara perquè està feta en negatiu enlloc d'en positiu), per exemple, enlloc d'especificar en quin sentit pensar és computar, dirà que computar no és la metàfora computacional. L'argument es podria anomenar "but that does not mean that":

> Thinking is computation, I claim, but that does not mean that the computer is a good metaphor for the mind. The mind is a set of modules, but the modules are not encapsulated boxes or circumscribed swatches'on the surface of the brain. The organization of our mental modules comes / from our genetic program, but that does not mean that there is a gene for every trait or that learning is less important than we used to think. The mind is an adaptation designed by natural selection, but that does not mean that everything we think, feel, and do is biologically adaptive. We evolved from apes, but that does not mean we have the same minds as apes. And the ultimate goal of natural selection is to propagate genes, but that does not mean that the ultimate goal of people is to propagate genes.[527]

Per tant, per entendre en quin sentit el computacionalisme no és la metàfora computacional i com això al mateix temps col·labora en la confecció d'una etologia digital, s'analitzaran cada una d'aquestes premisses afegint un símbol - que significa allò que no són, és a dir, la seva extensió en quant a minorització o retallada.

La premissa A-

La premissa A- defensa que la ment és el que fa el cervell (com mirar és el que fan els ulls), i el cervell processa informació, per tant, pensar és algun tipus de computació; ara bé, això no vol dir que la metàfora computacional, és a dir, que el cervell sigui una computadora, sigui una bona metàfora. Per Pinker, la metàfora computacional no és una bona metàfora perquè hi ha diferències entre el cervell i un ordinador:

> The computational theory of mind is not the same thing as the despised "computer metaphor." As many critics have pointed out, computers are serial, doing one thing at a time; brains are

526 *Ídem.*

527 *Ibídem*, pàgs. 23-24.

> parallel; doing millions of things at once. Computers are fast; brains are slow. Computer parts are reliable; brain parts are noisy. Computers have a limited number of connections; brains have trillions. Computers are assembled according to a blueprint; brains must assemble themselves. […] The claim is not that the brain is like commercially available computers. Rather, the claim is that brains and computers embody intelligence for some of the same reasons.[528]

Per tant, tant cervells com ordinadors encarnen intel·ligència per algun dels mateixos motius, en concret, tenen una mateixa base computacional, però el cervell no és com un ordinador (i menys un de comercial). Abans de seguir es poden fer algunes observacions: hi ha certa equivalència no justificada entre pensar i computar, i entre intel·ligència i informació (com a conjunt de dades). Així, Pinker està assumint que, en la mesura que pensar és a intel·ligència el que computar és a informació, llavors es pot dir que computar és intel·ligència, perquè tant computar com pensar comparteixen certa base comú. Ara bé, caldria primer definir amb precisió els següents termes: intel·ligència, pensar i encarnar (es dona per fet que "computar" i "informació" ja hi ha consens en la seva definició). Després, caldria demostrar les següents relacions:

1. Entre intel·ligència i pensar: sobre si pensar és una de les formes de demostrar intel·ligència, però no l'única. Defensar que a la intel·ligència només s'hi arriba pensant va en contra, no només de la desacreditada teoria de les intel·ligències múltiples de Horward Gardner (1983), sinó de qualsevol diferència que es faci amb el vocabulari de la ment (intuir, abstreure, raonar, intel·ligir). És a dir, per molt que els circuits neuronals que ara mateix sembla que s'activen quan es fa qualsevol d'aquestes accions siguin els mateixos (*Parieto-Frontal Integration Theory*[529]), a nivell de procés mental és obvi que es pot diferenciar entre raonar i intel·ligir. Per tant, si aquesta diferència aporta quelcom i no es vol perdre, o bé encara no s'ha trobat la forma d'expressar-la neuronalment o bé neuronalment tot s'expressa igual i no hi ha commensurabilitat. En altres paraules, el vocabulari *top-down* no ha de per què encaixar amb el vocabulari *bottom-up* o, si més no, no ha de per què encaixar d'una única manera, és a dir, universalment.

2. Entre pensar i computar: sobre si pensar és computar o tota idea que es pensi és computable, o, en altres paraules, sobre si computar és una de les formes de pensar, però no l'única. En la mesura que computar és aplicar algun algorisme sobre un conjunt de dades i que aquest

528 *Ibídem*, pàg. 26.

529 DUNCAN, John; ASSEM, Moataz; SHASHIDHARA, Sneha (2020). "Integrated Intelligence from Distributed Brain Activity" en *Trends Cogn Sci.*, 24 (2020), pàgs. 457-460. Citat en DEFELIPE, Javier (2022). *De Laetoli a la Luna: El insólito viaje del cerebro humano*, Barcelona, Crítica, 2022, pàg. 198.

algorisme cal expressar-lo en un llenguatge formal perquè sigui implementable, es pot entendre que computar és un tipus de matematització (per tant, també una part d'allò que és matematitzable, doncs caldria excloure tot allò no computable). Per tant, si pensar és computar, llavors tot el que es pugui pensar, ha de poder-se matematitzar a través de la computació. Aquesta hipòtesi només és factible si la matemàtica no és només un llenguatge per descriure curosament el món, sinó que el món està format per estructures matemàtiques o per estructures que compleixen lleis matemàtiques. La primera opció apel·la a un realisme matemàtic que ja s'ha vist que von Neumann assumeix. La segona opció no necessàriament és realista en aquest sentit, però sí que assumeix que la matemàtica és una eina de descobriment, i no d'invenció. En altres paraules, cap dels dos supòsits entenen la matemàtica només com un llenguatge humà, sinó que doten la matemàtica d'una objectivitat independent dels humans, és a dir, en llenguatge de la filosofia moderna, les veritats matemàtiques són vàlides en tot l'univers.

Contra aquesta idea s'ha revoltat una part important de la filosofia de la ciència del segle XX, des de Kuhn a Latour, però el projecte de la digitalització, del qual forma part el projecte IA, sembla bastant aliè a les seves idees (a excepció d'alguns autors afins a l'àrea, com Weizenbaum, i també d'aliens, com Roberto Casati). La crítica que fan contra la digitalització global aquests pocs autors es basa en diferents aspectes, un dels quals és la naturalesa social dels humans, que a través del llenguatge ens dota d'una capa cultural presumiblement independent de la merament natural (Weizenbaum ho resumeix amb una idea: «A computer might win at chess but that did not mean it could change a baby's diaper»[530]); i l'altra és el problema ètic que pot representar (Casati ho resumeix amb una crítica a la colonització digital: «"Es pot, per tant s'ha de". Si és possible que determinada cosa o activitat migri cap al digital, ha de migrar»[531]). En els apartats següents es desenvoluparan aquestes dues idees per mostrar com per construir una etologia digital, peça central del projecte IA, és necessari amagar aquestes crítiques i les seves conseqüents alternatives.

3. Entre computar i informació o sobre si es podria computar amb altres formats. Sembla que la computació digital és la més eficient que s'ha aconseguit, però això no treu que la

530 WEIZENBAUM, Josep y WENDT, Gunna (2006). *Islands in the Cyberstream. Seeking Havens of Reason in a Programmed Society*, Litwin Books, Sacramento, CA, 2015, pàg. 18.

531 CASATI, Roberto (2013). *Elogi del paper. contra el colonialisme digital*, Sant Cugat del Vallès, Pol·len Edicions, 2022, pàg. 14.

digitalització s'hagi basat en un format concret de dades, enlloc d'altres possibles formats o d'un format canviant. Per tant, se separa la informació del soroll en la mesura que els valors expressats són més o menys pròxims al 0 o a l'1. Tot allò que no és 0 o 1 simplement no és. En aquest sentit, es reconeix que una mesura universal de dada vàlida, és a dir, la tria d'un projecte computacional discret, tria derivada de l'anàlisi de l'eficiència electrònica, també ha suposat l'oblit d'allò analògic, no només en un sentit ètic, sinó també físic: allò que escapi, volgudament o inconscientment del procés d'informació, només pot ser considerat soroll de fons.

Aquesta crítica es pot expressar també de la següent manera: si allò que determina què és una dada vàlida és la possibilitat de captar-la digitalment, llavors s'estan excloent per defecte altres formats d'encapsular la realitat que podrien constituir nous tipus d'informació. Weizenbaum ho descriu amb un acudit:

> One dark night a policeman comes upon a drunk. The man is on his knees, obviously searching for something under a lamppost. He tells the officer that he is looking for his keys, which he says he lost 'over there,' pointing out into the darkness. The policeman asks him 'Why, if you lost the keys over there, are you looking for them under the streetlight?' The drunk answers, 'Because the light is so much better here.'[532]

En la versió més acadèmica ell mateix ho tradueix així:

> Science can proceed only by simplifying reality. The first step in the process of simplification is abstraction. And abstraction means leaving out of account all those empirical data which do not fit the particular conceptual framework within which science at the moment happens to be working, which, in other words, are not illuminated by the light of the particular lamp under which science happens to be looking for keys.[533]

Per tant, el problema de la neutralitat de la dada és també un problema general de la ciència, però en la mesura que la ciència no es limiti a ser digital, llavors sempre és possible que apareguin noves dades, és a dir, dades de format diferent. El problema rau quan una tecnologia limita què pot considerar-se una dada vàlida. En aquest cas, com diu Mitchell, tot semblen vectors per un investigador de LLM. Aquest problema, de fons el mateix que descriu Lady Lovelace a la nota G, no és altre que el de la creativitat: «The Analytical Engine has no pretensions whatever to originate anything. It can do whatever we know how

532 WEIZENBAUM, Joseph (1976). *Computer Power and Human Reason*, Londres, Penguin Books Ltd, New Ed, 1984, pàg. 127.

533 *Ídem.*

to order it to perform. It can follow analysis; but it has no power of anticipating any analytical relations or truths. Its province is to assist us in making available what we are already acquainted with»[534]. La rèplica a aquesta objecció utilitzant la IA Genereativa també s'ha hagut de matisar després de les diferents mostres de plagi que s'han detectat en alguns d'aquests motors[535].

4. Entre informació i coneixement (i aquest amb la intel·ligència). Sembla raonable que no hi ha coneixement sense informació (certa), però també és raonable que la informació (certa), per si sola, no és coneixement: la informació és condició necessària, però no suficient perquè hi hagi coneixement. Al mateix temps, tot i que Pinker no utilitzi aquí el terme "coneixement", caldria fer evident la relació d'aquest coneixement amb la intel·ligència, doncs també s'associa intel·ligència a actuacions que semblen més instintives que fruit d'un esforç cognitiu, i, fins i tot si es vol mantenir certa diferència conceptual entre aquests dos termes (cosa que no està clar que Pinker pretengui), s'hauria de determinar si el coneixement és part de la intel·ligència (condició necessària però no suficient), o coneixement i intel·ligència són dues maneres de dir el mateix.

En qualsevol cas, hi ha una sèrie de problemes que Pinker no analitza aquí i l'objectiu prioritari sembla defensar la seva proposta de quelcom que s'ha deixat enrere, la metàfora computacional: es parteix d'aquesta però cal diferenciar-s'hi. Al fer-ho, fa evident que no hi ha una commensurabilitat clara entre el vocabulari de la ment i el vocabulari del cervell, i que el vocabulari de la computació, en bona mesura basat en el vocabulari de la informació, és aplicat al cervell com si hi hagués una teoria de la computació comú, enlloc d'una teoria de la informació comuna. Quan es fa això, és a dir, quan s'aplica una teoria de la computació (pensada per a ordinadors) al cervell, es produeix una inversió de la metàfora computacional, i per això Pinker diu que la metàfora computacional ja no és adequada, no perquè sigui poc precisa, sinó perquè, estrictament parlant (és a dir, des del punt de vista de com es va fonamentar la metàfora computacional des de Wiener a von Neumann), és insuficient per poder afirmar el següent:

> When engineers first came to understand flight as they designed airplanes, it provided insight as to how birds fly, because principles of aerodynamics, like shape of an airfoil or the interplay of lift and drag, are applicable both to planes and to birds. That doesn't mean that the airplane is a

534 LOVELACE, Ada Augusta (1842). "Note G" en *Sketch of The Analytical Engine* (Invented by Charles Babbage). Consultat el 8 d'agost de 2024 a: https://history-computer.com/Library/Sketch%20of%20Engine.pdf

535 MARCUS, Gary; SOUTHEN, Reid (06.01.2024). "Generative AI Has a Visual Plagiarism Problem" en *IEEE Spectrum*. Consultat el 8 d'agost de 2024 a: https://spectrum.ieee.org/midjourney-copyright

> good model of the birds. Birds don't have propellers and headphone jacks and beverage service, for example. But by understanding the laws that allow any device to fly, one can understand how natural devices fly. The human mind is unlike a computer in countless ways, but the trick behind computation is the trick behind thought, representing states of the world, that is, recording information, and manipulating the information according to rules that mimic relations of truth and statistical probability that hold in the world.[536]

Per a Pinker, en la mesura que els avions han permès entendre part de l'aerodinàmica que també es pot aplicar en els ocells, els ordinadors poden fer entendre part de la computació que també es pot aplicar al cervell, sense tenir en compte que la recopilació d'informació que fa un ordinador i un cervell poden ser completament diferents (d'aquí el *frame problem* descrit primer des d'un punt de vista lògic per McCarthy i Hayes el 1969[537], però després reinterpretat, per exemple, per Dreyfuss el 2007[538]) i que els ordinadors poden arribar a un mateix resultat que els humans per camins diferents (el *shorcut problem* descrit per Mitchell[539]). A més a més, com s'ha vist abans tant en Kuhn, Weizenbaum o O'Neil, aquest argument denota una confusió entre teoria, model i realitat: els principis de l'aerodinàmica (teoria) configuren un model (el del vol) que s'aplica per fer volar avions i per ajudar a entendre com volen els ocells, però els ocells no són avions i el model, en tant que simplificació, segur que encaixa molt millor per als avions que per als ocells. Per tant, encara que es modifiquessin els principis de l'aerodinàmica i s'obtingués un model que permetés construir avions més ràpids, aquest model no serviria per descriure millor el vols dels ocells, perquè no volaran més ràpid. I seria absurd forçar els ocells per fer-los encaixar amb el model obtingut per aquests principis. Caldria elaborar uns principis de l'aerodinàmica específics pels ocells. Això Pinker sembla acceptar-ho pels ocells.

Tanmateix, això que sembla obvi pels ocells, no ho és pel cervell, i molt menys per a les activitats que es relacionen amb el cervell, com l'aprenentatge i la salut mental, és a dir, la pedagogia i la psicologia, camps on s'han proposat teories que apliquen uns principis de la

536 PINKER, Steven (10.01.1997). "Organs of Computation: A talk with Steven Pinker" en *Edge*. Consultat el 27 de juliol de 2024 a: https://www.edge.org/conversation/steven_pinker-organs-of-computation

537 MCCARTHY, John; HAYES, Patrick J. (1969). "Some Philosophical Problems from the Standpoint of Artificial Intelligence", Sect. 2.1, *Machine Intelligence* 4, ed. Donald Michie (Elsevier, 1969), pàg. 463 i següents. Consultat el 6 d'agost de 2024 a: http://www-formal.stanford.edu/jmc/mcchay69.pdf

538 DREYFUS, Hubert L. (2007). "Why Heideggerian AI Failed and how Fixing it would Require making it more Heideggerian" en Artificial Intelligence 171 (2007) 1137–1160. Consultat el 6 d'agost de 2024 a: https://www.sciencedirect.com/science/article/pii/S0004370207001452

539 MITCHELL, Melanie (26.04.2021). "Why AI is Harder Than We Think" en *arXiv*, pàg. 3. Consultat el 18 d'agost de 2023 a: https://arxiv.org/abs/2104.12871v1

computació pensats per a computadors al cervell, encaixin o no encaixin. En el següent apartat, se'n relacionaran alguns exemples.

La premissa B-

La premissa B- afirma que, tot i que la ment està organitzada en mòduls o òrgans mentals especialitzats, aquests mòduls no són caixes encapsulades o regions específiques a la superfície del cervell.

D'alguna forma, aquesta premissa reconeix que hi ha certa incommensurabilitat entre la ment i el cervell o, com a mínim, un desconeixement prou important de com funciona el cervell, doncs afirmar que la ment funciona amb mòduls, però el cervell no, i seguidament no descriure com es fan les connexions denota que, més que una premissa, és un hipòtesi pendent de validar.

Ara bé, per poder-la validar caldria muntar el trencaclosques o mapa sencer del que passa sinàpticament quan un humà té un pensament, un sentiment o una creença d'algun tipus. El problema principal d'aquesta camí és que per veure quines connexions sinàptiques s'estan produint, primer cal observar-les, i el nombre de connexions sinàptiques hipotèticament observables, si no es vol caure ja en algun tipus de priorització teòrica, es calcula que pot ser de l'ordre de cent bilions (10^{13})[540]. Per tant, resulta inevitable haver de triar, assumint que tota tria ja és algun tipus de simplificació. En el millor dels casos, això permetria saber que sempre que es produeixen X-Y-Z connexions sinàptiques, la majoria de persones estant llegint paraules, per exemple (es menysté aquí el problema que representa aquest "la majoria de persones").

El segon interrogant sol ser abordat tant per la psicologia conductista com per la filosofia de la ment (i també la del llenguatge) i sol seguir l'estratègia *top-down*: quan una persona diu sentir S, es produeix un comportament, i aquest comportament està relacionat amb les connexions sinàptiques X-Y-Z. Aquí el problema té un nivell de complexitat més, doncs mentre que assumir que al llegir paraules totes les persones ho fan igual, costa més d'acceptar que tothom estima, odia o tem de la mateixa manera (si més no, primer caldria demostrar que la conducta C està relacionada unívocament amb les expressions E del tipus "t'estimo", "t'odio", "et temo" o "et crec" que diu sentir una persona quan sent S).

Per tant, sembla coherent pensar que només podríem establir una teoria realment causal si es fusionessin ambdues aproximacions: si i només si es produeixen les connexions sinàptiques X-Y-Z, llavors se sent S, es té un comportament C i s'expressa E en veu alta. I, fins i tot llavors, caldria

540 GOLDEN, Rebecca (2013). "Mind-Boggling Numbers: Genetic Expression in the Human Brain" en *Science 2.0*, 15.05.2013. Consultat el 8 d'agost de 2022 a: https://www.science20.com/rebecca_goldin/mindboggling_numbers_genetic_expression_human_brain-109345

plantejar-se si, com tot element format per partícules, quan s'observa aquest comportament en qualsevol dels seus punts, aquest comportament és el mateix que quan no s'observa (quan es diu "t'estimo" és quan se sent que s'estima... i si no es diu, no se sent o se sent diferent? O si quan s'observen les sinapsis, es comporten igual que quan no s'observen o, com a mínim, si hi ha cert grau d'incertesa inherent a tota partícula). Evidentment, en el millor dels casos es tracta d'una demostració especialment complexa, com explica de Felipe en un text que ja parteix d'un supòsit que no es compleix i que sembla que està lluny de complir-se:

> No obstante, imaginemos que sabemos todo sobre los elementos del sistema nervioso, sus conexiones, sus propiedades fisiológicas, moleculares, etc. ¿Podremos entender entonces cómo la actividad mental genera la actividad cerebral? Aquí nos enfrentamos a otro gran problema, y es que prácticamente ignoramos cómo se pueden unir los distintos niveles de conocimiento (molecular, subcelular, neurona/sinapsis, circuito local, etc.) a las funciones cerebrales superiores. En otras palabras, 1) ¿cómo funcionan los distintos tipos de neuronas -con sus características moleculares, morfológicas y conexiones de entrada y salida dentro de una red local (en un nodo) en donde interactúan asimismo con diversos tipos de células gliales, que también son elementos fundamentales para procesar la información?; 2) ¿cómo se integra la información generada en este nodo en un sistema a mayor escala como el sistema motor, formado por múltiples nodos cuyos circuitos locales son a su vez distintos entre sí y, además, participan en otros sistemas como por ejemplo, el sensorial?; 3) ¿es esencial toda la información que entra y sale de un nodo en un momento dado o solo se utiliza una parte que activa o inhibe rutas principales para cada tipo de acción? Y si es así, ¿cómo se seleccionan esas rutas?; 4) ¿cómo aparece una respuesta cerebral compleja al final de este recorrido por el bosque neuronal?[541]

Per tant, des del punt de vista del sistema nerviós, sembla que la premissa B- està molt lluny de poder-se donar per vertadera.

La premissa C-

La premissa C- afirma que el funcionament bàsic dels mòduls de la premissa B- és lògic i aquest funcionament està especificat (i se suposa que determinat) pel programa genètic, però aquesta determinació no és exhaustiva entre cada gen i cada tret, cosa que delega certa importància a l'aprenentatge (adquisició de l'entorn).

541 DE FELIPE, Javier (2022). *De Laetoli a la Luna: El insólito viaje del cerebro humano*, Barcelona, Crítica, 2022, pàgs. 208-209.

L'expressió o metàfora "programa genètic" fou creada per Ernst Mayr, per una banda, i per François Jacob i Jacques Monod, per l'altra, el mateix any, el 1961, però de forma independent[542]. La seva disputa és un exemple de la dificultat de pensar sense biaixos: per una banda, des de Darwin cal pensar l'evolució mancada de sentit, sense un propòsit; tanmateix, un pensament teleològic sembla enquistat en els humans i impregna qualsevol discurs. La diferència que s'ha comentat en el capítol 1 entre teleologia i teleonomia és també la base de la metàfora del programa genètic i un intent de reconciliar la tensió no resolta que malviu dins de cada camp, tal i com ho explica el mateix Jacob pel que fa a la biologia: «For a long time the biologist has been consorting with teleology as with a woman without whom he can't live, but with whom he doesn't want to be seen in public. To this hidden relationship, the concept of program gives a legal status [Jacob 1973, p. 17]»[543]. De fet, aquesta tensió no resolta és la que anima tant Monod com Wiener, com ha estudiat Lily E. Kay, historiadora de la ciència especialitzada en biologia molecular[544]: «[...] Wiener prophesied a cybernetics of heredity by invoking the then-dominant view of the primacy of proteins. . . . As in all transmissions of messages, such a protein-based genetic transmission could be ultimately explained by information theory [Kay 2000, p. 86]»[545]. Aquest interès en poder traduir qualsevol llenguatge a un de comú que permeti aplicar allò après en un camp a un altre –la fantasia final seria una teoria del tot, però realment del tot, com el cristianisme o la ontologia digital de Konrad Zuse, teoria que també han defensat Stephen Wolfram o Max Tegmark entre d'altres personatges mencionats en capítols previs– és reconeixible en expressions com aquestes: «a mutation seems to be a bit of noise which gets incorporated into a message» o «If I could see heredity in terms of message and noise, I could get somewhere»[546]. El llenguatge de la informació seria la base comú ideal entre la matemàtica i la física; tanmateix, com recorda Sabine

542 PELUFFO, Alexandre E. (2015). "The "Genetic Program": Behind the Genesis of an Influential Metaphor" en *Genetics,* 200 (3), juliol de 2015, pàg. 685. Consultat el 8 d'agost de 2024 a: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4512536/pdf/685.pdf

543 JACOB, François (1973). *The Logic of Life: A History of Heredity*, Nova York, Pantheon Books. Citat en PELUFFO, Alexandre E. (2015). "The "Genetic Program": Behind the Genesis of an Influential Metaphor" en *Genetics,* 200 (3), juliol de 2015, pàg. 688. Consultat el 8 d'agost de 2024 a: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4512536/pdf/685.pdf

544 "Lily Kay, 53, life sciences historian" en *MIT News on campus and around the world*. Consultat el 8 d'agost de 2024 a: https://news.mit.edu/2001/kay-0110

545 KAY, Lily E. (2000). *Who Wrote the Book of Life? A History of the Genetic Code*, Redwood City Stanford University Press. Citat en PELUFFO, Alexandre E. (2015). "The "Genetic Program": Behind the Genesis of an Influential Metaphor" en *Genetics,* 200 (3), juliol de 2015, pàg. 689. Consultat el 8 d'agost de 2024 a: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4512536/pdf/685.pdf

546 *Ídem.*

Hossenfelder: «La física no es matemáticas. Es elegir las matemáticas»[547]. En el següent capítol es tornarà a recórrer a aquesta autora per descriure un dels requisits conceptuals de la inversió de la metàfora computacional, això és, la prevalença del llenguatge formal.

Pinker fa equilibris per evitar caure en postulats excessivament teleològics, que l'acabarien acostant excessivament a una teoria del disseny intel·ligent i desacreditant-lo científicament, però sense deixar de defensar les idees de la psicologia evolutiva. Així ho explica John Dupré, filòsof de la biologia:

> En momentos de reflexión, casi todo el mundo coincide en que las disposiciones humanas se desarrollan como resultado de la interacción entre los atributos biológicos del organismo y el entorno en el cual ese organismo se desarrolla. […] Pero a pesar de esas coincidencias reflexivas es común que los adversarios acusen a los psicólogos evolutivos de determinismo biológico, por sugerir a veces la idea de que la conducta aparece de manera independiente del entorno; y también es común que los psicólogos evolutivos acusen a sus críticos de respaldar el enfoque de la «pizarra en blanco» respecto de la mente humana, un enfoque según el cual las disposiciones humanas no están restringidas ni se ven afectadas de ninguna manera por la biología.[548]

Un dels postulats bàsics d'aquest corrent al qual se sol associar Pinker, la psicologia evolutiva, cau, segons Dupré, en la fal·làcia genocèntrica: només si estan codificats en els gens, les característiques adaptatives d'un organisme poden ser incorporades en un llinatge[549]. Tanmateix, aquesta premissa és falsa, ja que menysté l'efecte de l'entorn i la cultura en la naturalesa humana, i, concretament, en els gens:

> Las recientes investigaciones en el campo de la biología molecular desdicen de manera aún más directa algunos presupuestos centrales del dogma, que afirma que la información sólo se transmite desde el genoma y nunca hacia él. Ahora se sabe que existen mecanismos mediante los cuales la célula actúa sobre el genoma para afectar así las circunstancias en las que se expresan los genes.[550]

Dupré critica també com, des d'aquesta perspectiva, s'acaba defensant que certs comportaments, com la violència o el masclisme, són inherents als humans en la mesura que estan escrits en els gens i tenien una funció adaptativa en algun moment de la vida dels homínids. De fet,

547 HOSSENFELDER, Sabine (2018). *Perdidos en las matemáticas: cómo la belleza confunde a los físicos*, Barcelona, Editorial Planeta, 2019, pàg. 303.

548 DUPRÉ, John (2003). *El legado de Darwin. Qué significa hoy la evolución*, Madrid, Katz Editores, 2009, pàg. 121.

549 *Ibídem*, pàg. 123.

550 *Ibídem*, pàgs. 127-128.

Pinker defensa l'aplicació del que ell anomena enginyeria inversa (expressió provinent del camp de la informàtica) a la psicologia, en una mostra més d'aquesta inversió de la metàfora computacional: «I think the key to understanding the mind is to try to "reverse-engineer" it, to figure out what natural selection designed it to accomplish in the environment in which we evolved»[551]. Dupré critica que només hi ha un pas entre aquest "designed" i el disseny intel·ligent:

> Y resulta notable que los más fervorosos partidarios de considerar la evolución como fuente del saber contemporáneo sean los que más se esfuerzan por reintroducir el concepto del diseño. Pero por supuesto, el diseño no es más que una metáfora con respecto a los organismos, y parece ser una metáfora extremadamente peligrosa (tal vez debería denominarse la Peligrosa Metáfora de Dennett).[552]

La referència a Dennett no és gratuïta: Dupré critica que també des de posicions plenament materialistes com la de Dennett es pretengui donar sentit a la naturalesa: el sentit és cosa de la ment, no del cervell. Per això, també critica la posició de Richard Dawkins al parlar del gen egoista, no només com a títol llaminer, sinó com a posició argumentativa que prioritza el gen per sobre de l'organisme, la transmissió genètica per sobre de la cultural. A l'atribuir una semàntica, és a dir, un significat mental al gen, està convertint l'atzar de la selecció natural en un *blueprint* de connexions, expressió que també ja s'ha descrit en les primeres pàgines d'aquest capítol. És qüestió de temps que la idea de programa genètic passi de voler dir "pla d'activitats previstes" (*programme*) a "successió de passos ordenats" (*program*) –«In U.S. English, the "program" belongs to computers and the "programme" is a schedule, an agenda»[553]–, i així, el que surti més a compte per entendre aquest ordre, ironitza Dupré, sigui preguntar a aquell qui les ha ordenat: «Por cierto, si se pueden atribuir los orígenes de algo a un diseñador inteligente, las intenciones del diseñador son el mejor lugar donde buscar para lograr verdadera comprensión de la cosa en cuestión».[554]

En qualsevol cas, és obvi que Pinker es podia haver estalviat parlar de "programa genètic" i referir-se a genoma o genotip, però segurament li interessava traslladar cert sentit d'automatisme que hi ha en el concepte de codi genètic al genoma o al genotip, per això va haver de recórrer a una

551 PINKER, Steven (10.01.1997). "Organs of Computation: A talk with Steven Pinker" en *Edge*. Consultat el 27 de juliol de 2024 a: https://www.edge.org/conversation/steven_pinker-organs-of-computation

552 DUPRÉ, John (2003). *El legado de Darwin. Qué significa hoy la evolución*, Madrid, Katz Editores, 2009, pàg. 132.

553 PELUFFO, Alexandre E. (2015). "The "Genetic Program": Behind the Genesis of an Influential Metaphor" en *Genetics,* 200 (3), juliol de 2015, pàg. 688. Consultat el 8 d'agost de 2024 a: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4512536/pdf/685.pdf

554 DUPRÉ, John (2003). *El legado de Darwin. Qué significa hoy la evolución*, Madrid, Katz Editores, 2009, pàg. 132.

expressió dels anys 60 que el 1997 havia quedat o lleugerament desfasada o que tenia una funció interpretativa clara: dotar de sentit l'evolució.

La premissa D-

La premissa D- afirma que les operacions dels mòduls o òrgans mentals van ser modelades (valgui la redundància) per la selecció natural per resoldre els problemes de la caça i la recol·lecció propis de la major part de la història evolutiva dels humans, però aquest modelat també permet pensar, sentir i dir coses que no tenen res a veure amb el que podien pensar, sentir i dir els simis dels quals els humans han evolucionat. En altres paraules, els humans i els simis comparteixen estructura cerebral, però no necessàriament mental (es considera que aquesta darrera frase també forma part de la extensió de D- enlloc de E-, però tampoc hi hauria cap diferència si s'associa a la següent premissa).

Amb aquest afegitó, Pinker pretén superar una altra crítica habitual a la psicologia evolutiva, que explica així Dupré:

> ¿El conocimiento de la evolución no puede decirnos nada acerca de cómo somos? Deseo sugerir que, en el nivel de especificidad que pretende un proyecto como la psicología evolutiva, la respuesta es casi seguro que no. […] Con frecuencia se señala que nuestros genomas han resultado un 98,4% idénticos a los de los chimpancés. Se nos invita a concluir que somos, contrariamente a lo que indicaban nuestras inflacionadas expectativas, un 98,4% idénticos a los chimpancés. Pero si eso significa algo (cosa que dudo), seguramente es algo falso. [...] en la medida en que los genomas se encuentran entre los rasgos más invariables de los diferentes organismos, son en realidad el último lugar donde deberíamos esperar que se encontrara explicación de los rasgos más específicos de los organismos.[555]

Com s'ha vist abans, Pinker segueix fent equilibris entre dues tensions: la necessitat de defensar a ultrança l'atzar propi de la selecció natural (la manca de propòsit o objectiu) i la voluntat de trobar un sentit. Entre aquest dos pols, el de la natura i el de la cultura, la posició de Pinker va buscant matisos per donar valor a la part innata sense treure valor a la part cultural. Per exemple, per defensar la part innata afirma:

> Look at number of legs –it's an innate property of the human species that we have two legs as opposed to six like insects, or eight like spiders, or four like cats– so having two legs is innate. But if you now look at why some people have one leg, and some people have no legs, it's completely due to the environment: they lost a leg in an accident, or from a disease. So the two questions have to be distinguished. And what's true of legs is also true of the mind.[556]

555 *Ibídem*, pàg. 143.

Mentre que per defensar el pes de l'aprenentatge i l'entorn, afirma: «To believe that there's a rich innate structure common to every member of the species is different from saying the differences between people, or differences between groups, come from differences in innate structure»[557]. Tanmateix, no sembla ni que l'exemple de la cama acabi d'estar ben trobat –si segons Pinker, la ment és la computació del cervell, l'evolució de la cama és equivalent a les diferents formes de caminar?– ni que atorgui suficient pes a la dimensió cultural humana. Per una banda, l'evolució del cervell és dels aspectes més recents, com es veu entre les diferents famílies d'*Homo* i el canvi de dimensió d'aquest òrgan, mentre que l'evolució de les cames és dels més antics dins la família dels hominins, i s'especula si l'evolució de les cames (entre d'altres) no fou la causa de l'evolució posterior del cervell:

> In addition, the hypothesis also explains another important fact—namely, that the increase in brain size in human evolution occurred well after hominids became proficient hunters. The first individuals of H. erectus had brains between 600 cm3 and 900 cm3 , but, by a million years ago, H. erectus brains were larger than 1000 cm3 . It is reasonable to speculate that the evolution of endurance-running capabilities, which permitted persistence hunting, released a constraint on the size of the brain, a costly organ to grow and to maintain. In turn, bigger brains set the stage for the evolution of our own species, H. sapiens, in Africa sometime in the last 300,000 years.[558]

Per altra banda, l'evolució del cervell sembla haver de ser condició necessària per a l'evolució de la ment, però l'evolució de la ment o de les formes de pensar, sentir i fer, canvia a una velocitat molt diferent de l'evolució del cervell, o de caminar o de les cames. En qualsevol cas, hi ha quelcom que no acaba d'encaixar, com si Pinker forcés l'analogia.

Aquesta força aplicada a resoldre la tensió entre els dos pols sembla inherent a pensar la cultura com quelcom diferent de la natura, enlloc de pensar la cultura com una forma de pensar la natura. És a dir, és la mateix tensió irresoluble que hi ha entre la ment i el cervell (o cos i ànima) quan es prenen per separat i s'intenten, al mateix temps, integrar d'alguna forma. De fet, ja s'ha vist que Dennett era partidari d'una integració que, a la llarga, permetés explicar la ment en vocabulari del cervell, però que davant de l'envergadura i dificultat del problema, optava per proposar una teoria de la ment que complís amb els límits de la computació. Pinker fa quelcom similar, però

556 PINKER, Steven (10.01.1997). "Organs of Computation: A talk with Steven Pinker" en *Edge*. Consultat el 27 de juliol de 2024 a: https://www.edge.org/conversation/steven_pinker-organs-of-computation

557 *Ídem*.

558 LIEBERMAN, Daniel E. (2010). "Four Legs Good, Two Legs Fortuitous: Brains, Brawn, and the Evolution of Human Bipedalism" en *In the Light of Evolution* (Jonathan B Losos, ed.) Greenwood Village, CO: Roberts & Co, pàg. 16. Consultat el 9 d'agost de 2024 a: https://scholar.harvard.edu/files/dlieberman/files/2010g.pdf

enlloc dels límits de la computació, la computació ara és l'estructura comú, el que preval en l'encaix, i la resta, tant la ment com el cervell, s'hi han d'emmotllar.

La premissa E-

La premissa E- afirma que la tasca principal dels ancestres dels humans fou replicar-se genèticament (es pot suposar que les subtasques eren el caçar i recol·lectar de la premissa D), però això era i és una tasca inconscient i no un objectiu premeditat per part seva o nostra (dels humans actuals, s'entén).

A l'igual que Dawkins, Pinker defensa cert comportament "egoista" i, des del punt de vista de la ment, inconscient, dels gens, és a dir, de la natura, cosa que converteix a l'ésser humà, en part, en un titella dels seus òrgans: «Dawkins explained the theory in a book called *The Selfish Gene*, and the metaphor was chosen carefully. People don't selfishly spread their genes; genes selfishly spread themselves»[559]. D'alguna forma, cada gen trasllada aquest *conatus* a una estructura superior natural fins que es fa un salt a una estructura cultural, com la ment, però la ment, defensa Pinker, no ha de per què seguir aquesta "ordre" que li ve de les "entranyes" i de la qual no és conscient del tot, i pot ordenar al cos que actuï de forma diferent (diferent a què, si no és conscient de l'ordre, tampoc queda clar).

Aquesta concepció de l'ésser humà i la responsabilitat recorda l'aristotèlica: l'ésser humà és l'únic ésser viu que pot triar entre les seves potencialitats innates quina actualitzar, per tant, tot i tenir una tendència natural i uns límits genèticament marcats (en paraules d'ara), és responsable de les seves actuacions. Al mateix temps, l'ésser humà no pot negar la seva naturalesa i, com a animal racional, ha de trobar temps tant per alimentar-se com per pensar. Trobar un equilibri, un just terme mig, és la clau de la felicitat. Ara bé, aquesta concepció incloïa una visió plenament teleològica de la naturalesa la qual representa que, amb la selecció natural, ha quedat superada. Costa imaginar com sintetitzar novament aquestes dues pulsions, i l'explicació computacional sembla una bona solució, com anteriorment defensava Mayr de forma explícita:

> The term "final cause" goes back to Aristotle and means, from his own formulation, "in the purpose of what" something exists or takes place. For example, the adult individual is the purpose of why ontogenesis takes place. [...]The researchers working on teleology ended up discovering suitable concepts used in cybernetics and the information theory and adapted them well. The result was the development of a new language where appeared the words like "information", "program" and "retroactions". This language allows avoiding of traditional objections made against the teleological language [Mayr 1981, pp. 110–113].[560]

559 PINKER, Steven (1997). *How the mind works*, Londres, Penguin, 1998, pàg. 44.

Pinker sembla compartir aquesta idea i, d'alguna forma, les restriccions a les quals Dennett apel·lava per defensar la metàfora computacional (mecanicitat, finitud i temporalitat), no només són restriccions sinó que en Pinker impliquen assumir un nou vocabulari i, juntament aquest vocabulari, una nova forma d'entendre tant els cervell com la ment:

> On this view, psychology is engineering in reverse. In forward-engineering, one designs a machine to do something; in reverse-engineering, one figures out what a machine was designed to do. Reverse-engineering is what the boffins at Sony do when a new product is announced by Panasonic, or vice versa. They buy one, bring it back to the lab, take a screwdriver to it, and try to figure out what all the parts are for and how they combine to make the device work. We all engage in reverse-engineering when we face an interesting new gadget.[561]

Potser la metàfora computacional ha quedat enrere davant del computacionalisme, però és clar que el computacionalisme té davant la metàfora informàtica (o del programador, en tant que enginyer d'algoritmes), és a dir, aquella que defensa que així com un programador planifica i programa un algoritme, la naturalesa, en la mesura que se li pot aplicar la mateixa teoria matemàtica de la informació, ha d'haver seguit un curs similar. Per tant, des d'aquesta perspectiva, si s'aplica tot allò que s'ha après en el camp de la computació al cervell, encara que això impliqui forçar algunes semblances, ha d'acabar encaixant. Si s'aplica tot allò que s'ha après en el camp de la computació a la psicologia, ha d'acabar encaixant. Si s'aplica tot allò que s'ha après en el camp de la computació a l'educació, ha d'acabar encaixant. Perquè tot allò que s'ha après en el camp de computació diu que la ment funciona amb peces, com funciona un ordinador: amb una unitat central (el cervell), amb circuits lògics, amb peces de memòria temporal o d'execució i peces de memòria a llarg termini, amb perifèrics que capten *inputs* i perifèrics que transmeten *outputs*.

Aquesta és la ment que cal trobar en el cervell: perquè la tesi de que la ment és el que fa el cervell només vol dir que, com que els humans pensem en informació, el cervell ha de pensar en informació (premissa A^{-1}); perquè com que els humans imaginem peces o mòduls, el cervell ha de tenir peces o mòduls (premissa B^{-1}); perquè com que els humans hem inventat la lògica, el cervell ha de funcionar lògicament (premissa C^{-1}); perquè com que els humans cacem i recol·lectem, el cervell ha de contenir els motius o cromosomes d'aquest comportament (premissa D^{-1}); perquè com que els humans jutgem l'egoisme, els gens s'hi han de comportar (premissa E^{-1}).

560 MAYR, Ernst (1981). *La biologie de l'évolution.* Citat en PELUFFO, Alexandre E. (2015). "The "Genetic Program": Behind the Genesis of an Influential Metaphor" en *Genetics,* 200 (3), juliol de 2015, pàg. 687. Consultat el 8 d'agost de 2024 a: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4512536/pdf/685.pdf

561 PINKER, Steven (1997). *How the mind works*, Londres, Penguin, 1998, pàg. 21.

És la ment la que ha modificat el cervell, és la idea la que ha canviat la matèria. El computacionalisme és, tot i l'aparent constructe mecanicista, un mena d'idealisme.

5.3 Recopilació de trets de la metàfora computacional

A mode de resum del capítol, es recopilen a continuació algunes de les premisses que han anat sorgint durant l'anàlisi de la metàfora computacional, especialment, aquelles que sembla que tenen alguna influència en la confecció d'una etologia digital.

1. En Black, s'ha vist que la *interaction view* defensa que hi ha una influència bidireccional entre el primer terme de la metàfora i el segon. Aquí s'ha descartat que aquesta influència sigui recíproca, però sí que s'ha acceptat que, com a norma general, és el primer terme el que rep el significat d'una simplificació del segon terme. Així, "l'home és un llop" significa que uns atributs comunament associats als llops s'assignen als humans. Aquests atributs són una simplificació basada en unes idees preestablertes en l'imaginari col·lectiu, altrament s'ha d'explicar en quin sentit es diu i la metàfora no funciona per si sola.

2. En Boyd, s'ha vist que l'obertura pròpia d'una metàfora quan aquesta té una funció en la ciència, també pot tancar-se i petrificar-se, cosa que pot provocar una interpretació literal de la mateixa. Així, un forat negre és un forat, mentre que la pota d'una taula no és una pota (o no és la pota d'un animal, però també és una pota). En qualsevol cas, sembla que en l'ús de metàfores sempre hi ha el perill de caure en una interpretació literal, especialment si qui la fa servir desconeix el camp en el qual ha sorgit la metàfora.

3. En Kuhn, s'ha plantejat la diferència entre les idees, les paraules (entre elles, les metàfores) i les coses, i s'ha situat el model com algun tipus de figura intermèdia entre les idees i la realitat, entre les fórmules matemàtiques i la naturalesa. El paper dels models i el llenguatge formal amb el qual se solen confeccionar també poden explicar com algunes metàfores tenen més o menys èxit (cal recordar la teoria de C.C. Anderson sobre la necessitat que la matemàtica aplicada al camp d'origen es pugui aplicar al camp de destí).

4. En Wiener, McCulloch, Shannon i von Neumann, és a dir, els artífexs o creadors de la metàfora computacional, s'ha vist que la comparació entre l'ordinador i el cervell (en aquest ordre) partia del reconeixement d'una ignorància, especialment del cervell, però secundàriament també de l'ordinador. Aquesta ignorància es pal·liava amb una simplificació.

5. També s'ha vist com la simplificació del cervell es feia per propiciar l'equivalència amb l'ordinador a través de propietats elèctriques, excloent explícitament propietats químiques i mecàniques de les sinapsis.

6. També és propi d'aquests primers autors una certa cura en fer les afirmacions: la metàfora computacional no es dona per vàlida encara, sinó que s'ha de demostrar la seva validesa.

7. Aquests autors també comparteixen cert desinterès per la ment: el tema és el cervell i com replicar alguna part del seu funcionament en ordinadors. Així, el sentit de la metàfora és del cervell cap a l'ordinador, en concret, uns atributs comunament associats al cervell (com pensar), s'assignen a l'ordinador. Recuperant de nou el paral·lelisme amb la metàfora política, "L'home és un llop", es pot escriure la metàfora computacional com "L'ordinador és [com] un cervell".

8. En Shannon, s'ha vist com el vocabulari imperant per descriure la comparació del cervell amb l'ordinador serà el de la teoria de la informació, que servirà de pont entre el vocabulari de l'ordinador i el vocabulari del cervell.

9. En von Neumann, s'ha vist que opera un cert realisme matemàtic pel qual l'ordre i connexió de les idees és l'ordre i connexió del món o, com a mínim, es prioritza la matemàtica com a llenguatge per descriure objectivament el món.

10. En Weizenbaum, s'ha vist que la diferència entre el mode performatiu i el mode teòric no sempre es té en compte quan s'implementa computacionalment una teoria sobre la naturalesa i que això pot produir certa confusió. En concret, es pot pensar que una bona solució performativa, és a dir, informàtica, pot ser una bona solució teòrica. En el cas d'una simulació, això encara és més complicat de determinar perquè una simulació pot ser performativament millor sense que això impliqui cap millora en la comprensió del fenomen natural estudiat.

11. També s'ha vist el fort paper manipulatiu que té la informàtica per representar: fins i tot sabent que un programa és un conjunt d'algoritmes, un usuari final pot tenir la sensació d'estar tractant amb una entitat viva (com exemplifica l'anècdota de la secretària de Weizenbaum).

12. També s'ha vist la preocupació de Weizenbaum (i de O'Neil) per com de fàcil els humans deleguem problemes en la tecnologia, fins i tot quan aquests problemes a encarar són de caràcter social i no tecnològic. Aquesta delegació, tant de solucions com també de preguntes, és també una delegació de responsabilitats.

13. En Dennett, s'ha vist com quan la metàfora computacional s'assenta en un camp, en el seu cas el de la filosofia (però també en el de la psicologia), el llenguatge d'aquell camp es comença a "contaminar" del llenguatge que introdueix la metàfora (aquesta característica ja l'havia mencionat també Boyd). En aquest cas, comença a no estar clar si la comparació és entre el cervell i l'ordinador o entre la ment i l'ordinador. No es tracta que Dennett ho confongui, sinó que posa sobre la taula que hi ha diferents propostes sobre el tema: la dualista (que diferencia ment de cervell), l'emergentista (que mistifica el sorgiment de la ment), la neurocièntífica (que busca una agulla en un paller, ja sigui *top-down* o *bottom-up*) i la idealització teorètica de la ment (que proposa una actualització del problema clàssic cos-ànima amb les restriccions pròpies de la computació, això és, mecanicitat, finitud i temporalitat). La seva proposta, la idealització de teorètica de la ment és paradoxal, com un camí que no porta enlloc.

14. En Pinker, s'ha vist com la metàfora computacional, originalment del cervell cap a l'ordinador, s'inverteix: l'ordinador ja no és com un cervell, sinó que el cervell és com un ordinador. Aquesta inversió implica que qui s'ha d'emmotllar per encaixar a uns atributs que no li són propis (ja que formen part de l'imaginari col·lectiu de l'altre terme) és el cervell.

15. També s'ha vist que aquesta inversió es produeix a causa d'una sèrie de reduccions entre intel·ligència i pensar, entre pensar i computar, i entre computar i dada (*bit*). Així, el bit (base de la teoria de la informació, per tant, d'una teoria sobre quelcom no material sinó ideal en un sentit modern) condiciona com computar i, en la mesura que computar és l'única forma de pensar, propietat normalment associada al cervell, condiciona també la matèria.

16. Per tant, el vocabulari de la informació adquirit pel llenguatge de la ment, és el que s'utilitza per descriure el cervell, cosa que obliga a fer una sèrie de matisos. En altres paraules, defensa una unitat de mesura entre ment i cervell (en aquest ordre) basada en la informació.

17. Finalment, també en Pinker però explícitament en Mayr, s'ha vist que el problema de fons és una tensió entre una concepció de la naturalesa basada en l'atzar i una concepció de la vida (en tant que estudi de la biologia) basada en el sentit. La reconciliació d'aquestes dues

visions o, com a mínim, un intent d'equidistància o equilibri entre ambdues tensions, es veu possible si s'aplica la teoria de la informació operant en el camp de la computació, a la psicologia i, per extensió, a la biologia.

18. Hi ha autors, com Weizenbaum o Casati, que defensen que una altra manera d'entendre la digitalització és possible (a part de més humana i més ètica).

En el següent capítol, s'analitzarà com aquestes premisses han fonamentat conceptualment una etologia digital.

Thus, “entropy never can decrease” means
exactly the same as “information never
can increase.”

Warren S. McCulloch, “The brain as...”

6. Etologia i metàfora computacional

En els capítols 2, 3 i 4, s’han agrupat una sèrie de proposicions, no només del món divulgatiu sinó també i especialment del món acadèmic, en dos grans blocs: un bloc de textos que utilitza l’estratègia de la por (capítol 2), i un bloc de textos més escèptics i que sovint alerten contra l’estratègia de la por (capítol 4). Entremig, s’han situat dos textos frontissa, la carta de Musk i la carta de Gates, que mostren com es pot defensar el mateix fent ús del mateix format (la carta oberta, divulgativa), però seguint aquestes dues estratègies diferents: Musk, la de la por; Gates, la de la il·lusió. És a dir, el problema no és el públic al qual va adreçat un discurs o un altre, sinó el respecte per aquest públic. Secundàriament, també s’ha defensat que l’estratègia escèptica, tot i tenir menys altaveus i ser menys cridanera, a llarg termini pot ser més fructífera, cosa que la feia més perillosa en vistes a la confecció d’una etologia digital.

En els següents tres apartats, es classificaran les proposicions analitzades en els capítols 2, 3 i 4 per veure si hi ha tendències comuns dins de cada estratègia dependents del mecanisme utilitzat.

6.1 Les diferents estratègies i mecanismes per construir una etologia digital

Al primer capítol d’aquest treball, s’ha diferenciat entre una aproximació intuïtiva a una etologia digital i una aproximació raonada. La primera era resultat d’una reacció espontània, no elaborada, de la lectura d’una sèrie de titulars de diaris o comentaris divulgatius d’especialistes. La segona, en canvi, era el resultat d’un estudi dels fonaments de l’etologia tal i com els explicava i enumerava un dels seus fundadors, Konrad Lorenz.

A continuació es farà una compilació d’aquestes 182 proposicions agrupades per estratègies (intuïtiva o raonada) i mecanismes (de A-I, mecanismes propis d’una estratègia intuïtiva; de 1-10, mecanismes propis d’una estratègia raonada). Consideracions prèvies: tal i com s’explica a l’annex 1, on hi ha la relació de totes les proposicions dins de cada categoria, és discutible si una proposició és d’un tipus o d’un altre o d’ambdós. En tots els casos menys un, s’ha assignat cada proposició a una categoria o una altra (l’excepció és P23, que té elements molt clars de dues categories i es fa impossible de decidir-se). També cal tenir en compte que algunes proposicions formen part d’un mateix argument elaborat per parts: en aquest cas, només compten una vegada. Per tots aquests

motius, el nombre total de proposicions, 182, no encaixa amb el sumatori de les parts, si no es tenen en compte aquestes consideracions. Quan ha semblat que podia ser útil, s'han acompanyat les dades amb una gràfica. Aquesta gràfica no té cap pretensió estadística: és un recurs explicatiu més.

- Proposicions d'autors associats a l'estratègia de la por: Russell, Bostrom, Chalmers i Musk: 90
  - Russell: 31
  - Bostrom: 17
  - Chalmers: 14
  - Musk: 28
- Proposicions d'autors associats a l'estratègia de l'escepticisme: Gates, Marcus, Mitchell i Brooks: 92
  - Gates:23
  - Marcus: 12
  - Mitchell: 25
  - Brooks: 32

Desglòs de resultats en l'aproximació intuïtiva

• Total de proposicions: 115

A) Ús antropomòrfic de termes aplicant-los a un ens digital (i.e., l'ordinador està pensant):

1. Total de proposicions: 20
2. Distribució segons estratègia:
   - Autors de la por: 11
   - Autors escèptics: 9
3. Detall:

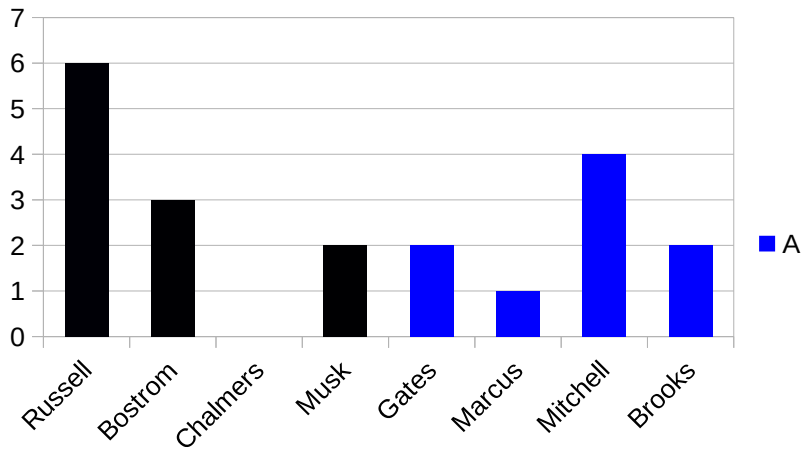

Figura 12: Mecanisme A

B) Ús de metàfores o comparacions entre elements analògics i digitals (i.e., l'ordinador funciona com un cervell):

1. Total de proposicions: 17
2. Distribució segons estratègia:
   1. Autors de la por: 6
   2. Autors escèptics: 11
3. Detall:

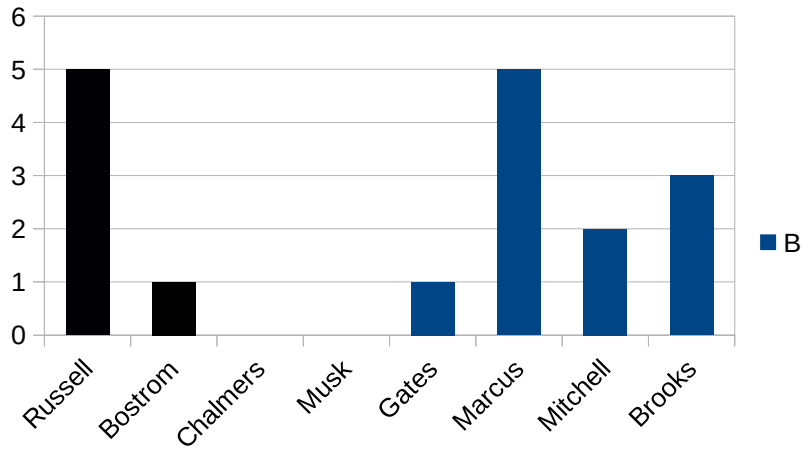

Figura 13: Mecanisme B

C) Descriure el comportament d'un ésser viu amb vocabulari propi d'un ens digital (i.e., els nens són *multitasking*):

1. Total de proposicions: 7
2. Distribució segons estratègia:
   1. Autors de la por: 3
   2. Autors escèptics: 4
3. Detall:

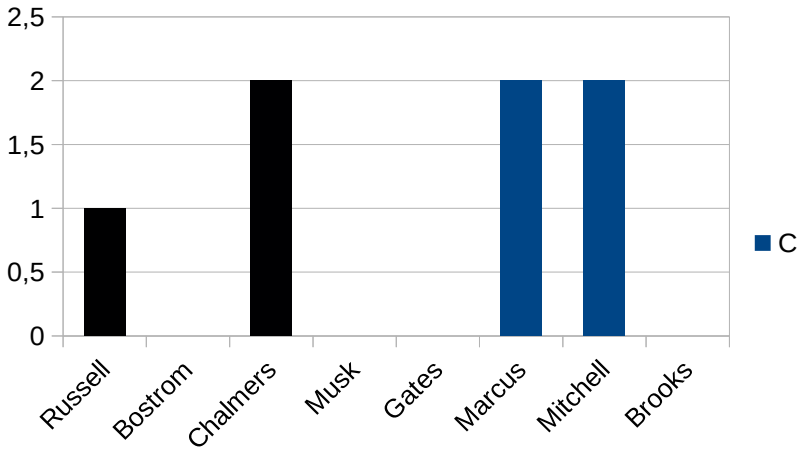

Figura 14: Mecanisme C

D) Disminuir capacitats o atributs d'uns ésser viu per tal que un ens digital s'hi pugui comparar (i.e. els humans tampoc som tan intel·ligents com ens pensem).

1. Total de proposicions: 6
2. Distribució segons estratègia:
   1. Autors de la por: 1
   2. Autors escèptics: 5
3. Detall:

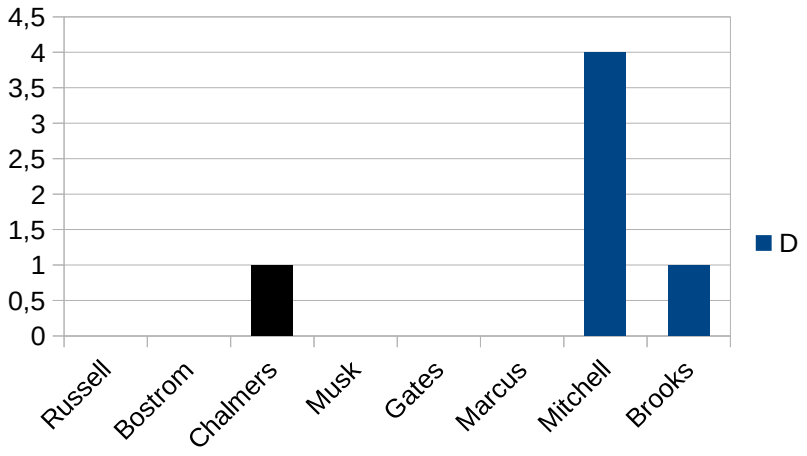

Figura 15: Mecanisme D

E) Desvincular l'eina respecte el seu creador, però això no representa necessàriament una amenaça (i.e. l'aparició de la IA):

1. Total de proposicions: 24
2. Distribució segons estratègia:
   1. Autors de la por: 17
   2. Autors escèptics: 7
3. Detall:

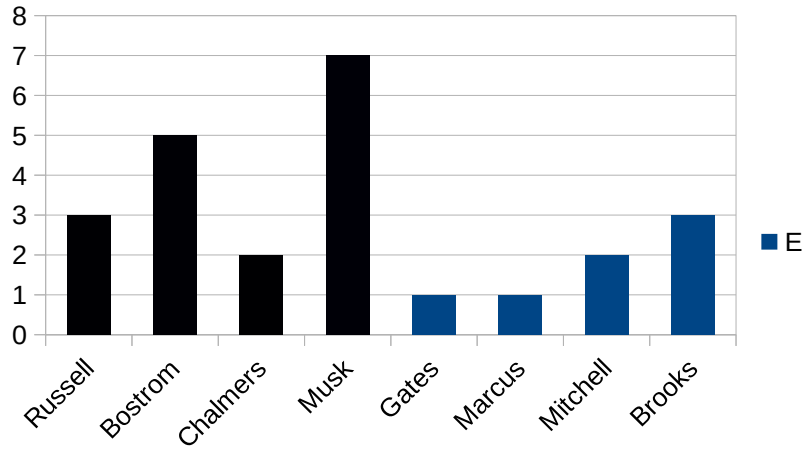

Figura 16: Mecanisme E

F) Apel·lar a la inexplicabilitat del seu funcionament (i.e., no sabem com ho fan):

1. Total de proposicions: 4
2. Distribució segons estratègia:
   1. Autors de la por: 2
   2. Autors escèptics: 2
3. Detall:

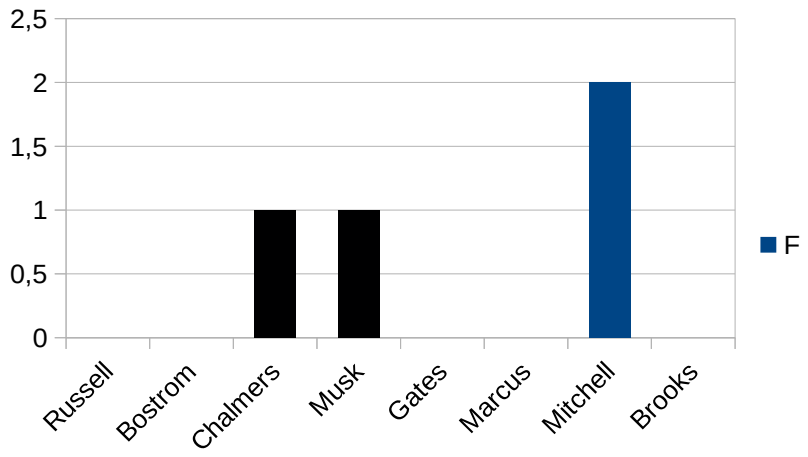

Figura 17: Mecanisme F

G) Assumir que els humans, de fet, som màquines (i.e., el cervell literalment és un computador):

1. Total de proposicions: 0

H) Presentar la IA com la nova pedra filosofal que resoldrà els problemes de la humanitat (i.e., la IA resoldrà el canvi climàtic):

1. Total de proposicions: 22
2. Distribució segons estratègia:
   1. Autors de la por: 3
   2. Autors escèptics: 19
3. Detall:

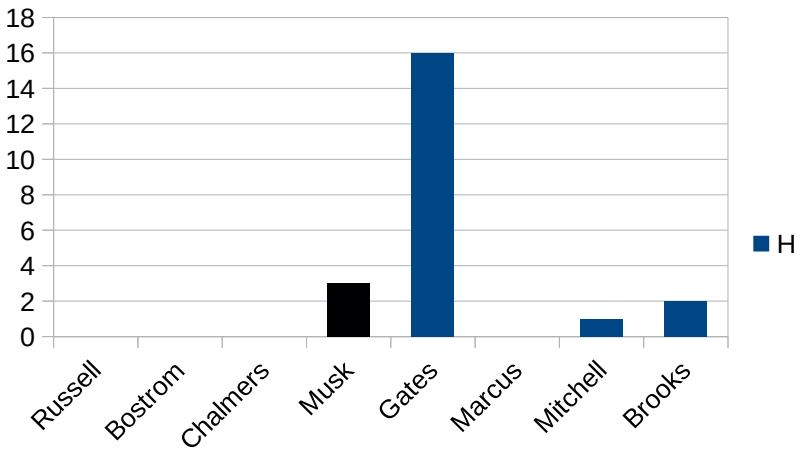

Figura 18: Mecanisme H

I) **Presentar la IA com una nova espècie invasora** amb la qual cal negociar i entendre-s'hi, ja que és una amenaça (i.e., els humans estem en risc d'extinció a causa de la IA):

1. Total de proposicions: 15
2. Distribució segons estratègia:
   1. Autors de la por: 15
   2. Autors escèptics: 0
3. Detall:

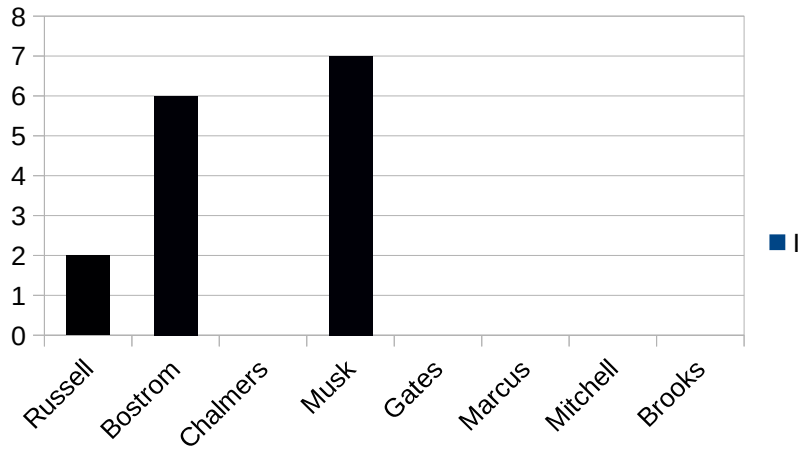

Figura 19: Mecanisme I

Desglòs de resultats en l'aproximació raonada[562]

• Total proposicions: 13

1. La pretensió innecessària de totalitat.
   1. Total de proposicions: 2
   2. Distribució segons estratègia:
      1. Autors de la por: 1 (Musk)
      2. Autors escèptics: 1 (Mitchell)
2. La pretensió que aquesta totalitat té un sentit o direcció predefinida.
   1. Total de proposicions: 1
   2. Distribució segons estratègia:
      1. Autors de la por: 0
      2. Autors escèptics: 1 (Marcus)
3. La pretensió que aquest sentit és natural.
   1. Total de proposicions: 2
   2. Distribució segons estratègia:
      1. Autors de la por: 1 (Russell)

562 Només es mostren gràfics quan això ajuda a la comprensió dels resultats

2. Autors escèptics: 1 (Marcus)

4. La pretensió que aquesta naturalitat converteix en una afició innòcua el tracte amb les entitats digitals.
   1. Total de proposicions: 0

5. La pretensió que un programa ha de ser alliberat per gaudir de la seva màxima expressió.
   1. Total de proposicions: 0

6. La pretensió que l'actuació d'un programa alliberat permet extreure conclusions que sobrepassen al propi programa, com la de si hi ha trets adquirits o innats.
   1. Total de proposicions: 0

7. La pretensió que un programa informàtic pot tenir comportaments patològics i inesperats.
   1. Total de proposicions: 1
   2. Distribució segons estratègia:
      1. Autors de la por: 1 (Chalmers)
      2. Autors escèptics: 0

8. La pretensió que els comportaments patològics d'un programa agafen per sorpresa al seu programador.
   1. Total de proposicions: 0

9. La pretensió que el programador fa una feina similar a una deïtat, ja que crea entitats.
   1. Total de proposicions: 0

10. La pretensió que el programador ha de seguir els passos de l'evolució per aconseguir que el seu programa pugui inscriure's coherentment a la naturalesa.
    1. Total de proposicions: 7
    2. Distribució segons estratègia:
       1. Autors de la por: 5
       2. Autors escèptics: 2
    3. Detall:

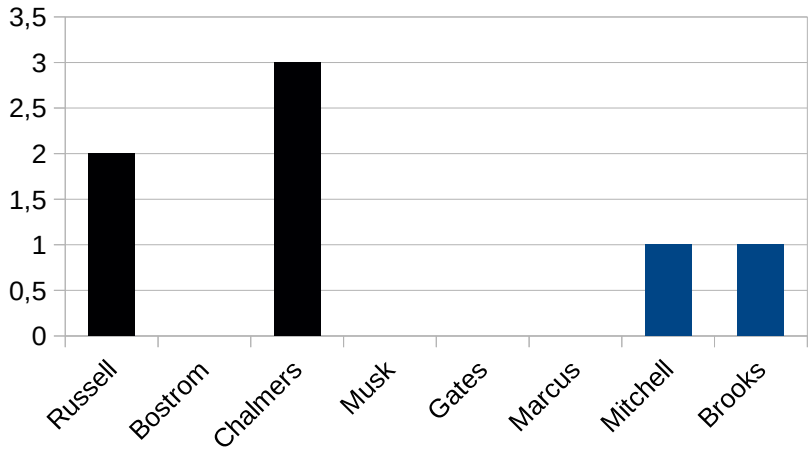

Figura 20: Mecanisme 10

Desglòs de resultats sense connotació etològica

1. Total de proposicions: 41
2. Distribució segons estratègia:
    1. Autors de la por: 13
    2. Autors escèptics: 28
3. Detall:

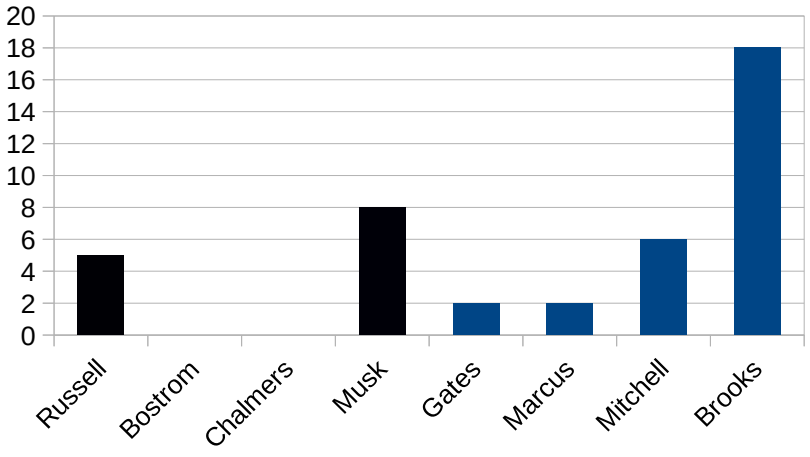

Figura 21: Proposicions sense connotació etològica

Anàlisi de les dades

La primera evidència és que les estratègies que intuïtivament permeten confeccionar una etologia digital són molt més comuns que les estratègies més raonades (115-13), i entre les intuïtives, els mecanismes més habituals són, en aquest ordre, els següents cinc: E (desvincular l'eina respecte el seu creador), H (presentar la IA com la nova pedra filosofal), A (ús antropomòrfic), i B (ús de metàfores i comparacions), i I (l'amenaça d'una nova espècie invasora).

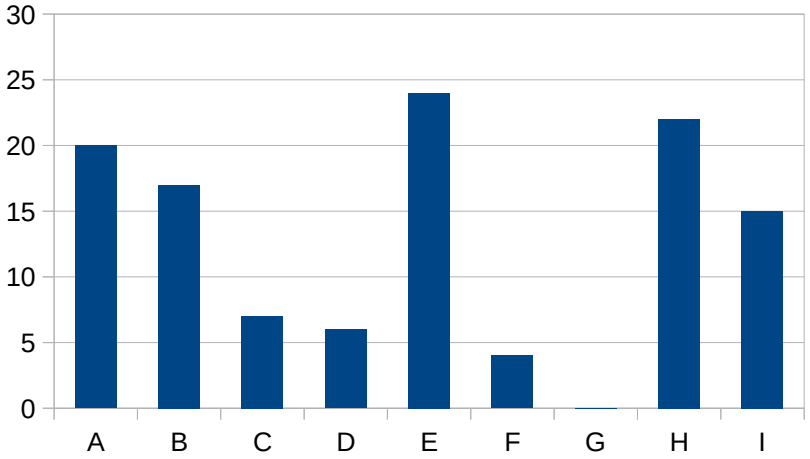

Figura 22: Comparació ús mecanismes intuïtius

En canvi, hi ha quatre mecanismes que, tot i que podrien ser intuïtivament fàcils de fer servir per confeccionar una etologia digital, no sembla que, si més no en aquest textos, hagin estat preferents. De fet, un mecanisme com el G (assumir que els humans són màquines), que s'ha vist utilitzat en autors que es consideren pares de la metàfora computacional, com Shannon, Minsky, Simmons o Nilson, no s'ha detectat que sigui un mecanisme utilitzat més recentment. També es constata que els autors de la por són els únics en utilitzar una estratègia com la I (l'amenaça d'una nova espècie invasora), i que Bill Gates és el més optimista de tot el grup (dels 22 casos d'ús del mecanisme H, el de la pedra filosofal, Gates n'és responsable de 16). D'alguna forma, els clixés se solen complir.

En general, els autors associats a l'estratègia de la por, opten per mecanismes més alienants, com l'E (desvincular l'eina respecte el seu creador) i l'I (l'amenaça d'una nova espècie invasora), ambdós mecanismes vinculats amb el rebuig de l'altre, mentre que els autors associats a l'estratègia escèptica, opten principalment pel mecanisme H (presentar la IA com la nova pedra filosofal), cosa que converteix aquests escèptics (en el discurs de la por) en optimistes crítics.

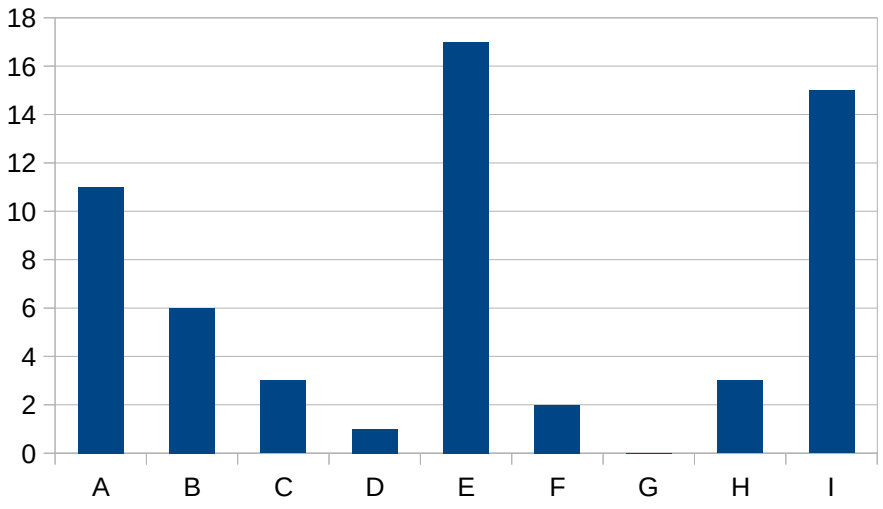

Figura 23: Mecanismes intuïtius utilitzats pels autors de la por

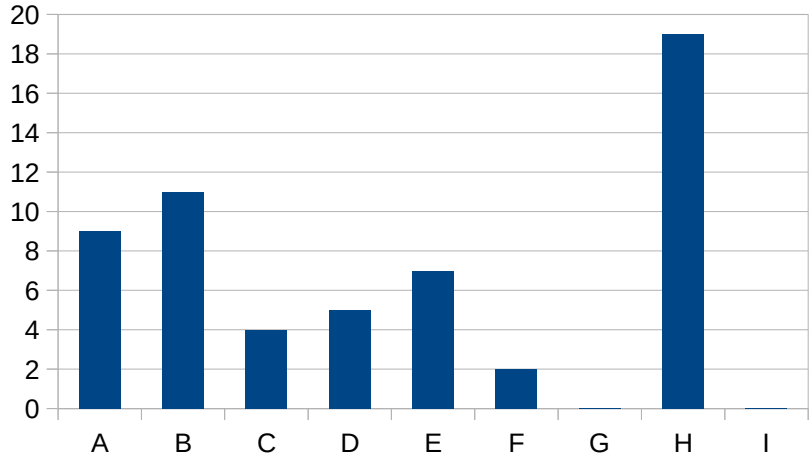

Figura 24: Mecanismes intuïtius pels autors escèptics

Tanmateix, hi ha dos altres mecanismes que també destaquen per ser utilitzats per autors d'ambdós sectors: l'A (ús antropomòrfic) i el B (ús de metàfores i comparacions). De fet, la diferència entre un mecanisme i l'altre és més interpretativa que real, doncs un ús antropomòrfic és un ús metafòric especialitzat en humans. Per exemple, la proposició P2: «Like any rational entity, the algorithm learns how to modify the state of its environment – in this case, the user's mind– in order to maximize its own reward», podria haver estat considerada metafòrica (que un algoritme aprengui és una metàfora antropomòrfica, doncs es diu en sentit figurat i fa una activitat que habitualment s'associa amb els humans). A més, cal considerar que les metàfores i comparacions que es fan servir en el camp de la IA són, per definició, amb els éssers humans, doncs són el model d'intel·ligència que tenen en ment des de la inauguració oficial del projecte a Dartmouth el 1956. Només en el cas que la metàfora o comparació es fes amb un altre animal, com quan Brooks es va plantejar començar per una formiga o un ratolí, es podria considerar que no és un ús antropomòrfic, però en aquest cas, aquest animal només seria un pas previ per aconseguir acabar creant una IA emulable a un ésser humà. Així, es proposa unificar ambdues estratègies seguint una pauta de conversió com la que s'aplica, a mode d'exemple, en les següents proposicions:

Proposició original	**Proposició final**
P26:«after seeing Arthur Samuel's checker-playing program learn to play checkers far better than its creator»	P26':«after seeing Arthur Samuel's checker-playing program **is like a student** that learn to play checker far better than its **teacher**»

P96:«AIs also make factual mistakes and experience hallucinations»	P96':«AIs mistakes are **like humans** factual mistakes and hallucinations»
P147:«their training in predicting words in vast collections of text has taught them the form of language but not the meaning».	P147':«their **optimization**, **due to** vast collections of text, is **how they can replicate** the form of language but not the meaning»
P179:«What the large language models are good at is saying what an answer should sound like, which is different from what an answer should be».	P179':«What the large language models are **useful for** is saying what an answer should sound like, which is different from what an answer should be»

Sembla que no seria especialment complicat, amb una mica de gràcia i perícia, desvelar la comparació que qualsevol proposició antropomòrfica implica i fer-la explícita. També és obvi que el mecanisme G (assumir que els humans són màquines) és una comparació sense el com, és a dir, una metàfora.

Aquesta mateixa conversió s'hauria de poder fer amb els altres tipus de mecanismes de forma més o menys fàcil, especialment amb l'E (desvincular l'eina respecte el seu creador), ja que la desvinculació d'un ens digital del seu creador no deixa de ser un ús antropomòrfic afegit a una omissió (la del creador) i, com s'ha vist abans, aquesta es pot transformar en una comparació i l'explicitació d'aquest creador:

Proposició original	**Proposició final**
P72:«sophisticated agents attempt to manipulate or directly control their reward signals»	P72':«**as if** sophisticated agents **that a human has programmed** attempt to manipulate or directly control the reward signals **that a human has previously configured**»
P77:«The emergence of artificial intelligence (AI) promises dramatic changes in our economic and social structures as well as everyday life in Europe and elsewhere; it has been compared to both electricity and the internet».	P77':«The **increasing use by humans** of artificial intelligence (AI) promises dramatic changes in our economic and social structures as well as everyday life in Europe and elsewhere; it has been compared to both electricity and the internet»

P116:«AI doesn't have to want to destroy us in order to create havoc. In the short term, what we should worry most about is whether machines are actually capable of reliably doing the tasks that we assign them to do»	P116':«**The programmers of** AI doesn't have to want to destroy us in order to create havoc. In the short term, what we should worry most about is whether **we can program machines that can** actually be reliably doing the tasks that we assign them to do»

També el mecanisme D (disminuir capacitats dels humans) és fàcilment tractable com a una comparació, ja que aquesta disminució és l'ocultació d'una expressió d'aquest tipus: "com si els humans només" o equivalent. És a dir, és una comparació en què s'amaga una simplificació:

Proposició original	**Proposició final**
P137:«We will use the umbrella term abstraction to denote this cluster of operations: seeing similarity (analogy-making), forming a concept, and applying a concept»	P137':«We will use the umbrella term abstraction **as if human abstraction was only that**, to denote this cluster of operations: seeing similarity (analogy-making), forming a concept, and applying a concept»
P164:«Building models that are below some complexity threshold also would mean that there is nothing in principle that we do not understand about intelligent or living systems».	P164':«Building models that are below some complexity threshold also would mean that there is nothing in principle that we do not understand about intelligent or living systems, **as if we could understand everything**»

El mecanisme F (apel·lar a la inexplicabilitat) es converteix en una comparació si s'explica allò que s'ha evitat explicar i se substitueixen les expressions misterioses per altres de més clares:

Proposició original	Proposició final
P50:«But as many people have observed, two decades ago, if we'd seen a system behaving as LLMs do without knowing how it worked, we'd have taken this behavior as fairly strong evidence for intelligence and consciousness»	P50':«But as many people have observed, two decades ago, if we'd seen a system behaving as LLMs do without knowing how it worked, we'd have taken this behavior as fairly strong evidence for intelligence and consciousness, **but, fortunately, we do know they are programmed by humans**»
P145:«How LLMs perform these feats remains mysterious for lay people and scientists alike»	P145':«How LLMs perform these **tasks is due to the fact that some humans have been working on deep learning for more than 50 years**, so that some scientists (specialists in LLM) could explain it to lay people»

I, per altra banda, els mecanismes H i I són actitudinals: denoten com es percep un canvi per part de qui el percep. Així, les proposicions de la categoria H denoten optimisme davant d'un invent, mentre que les de tipus I denoten pessimisme (i dramatisme).

L'únic mecanisme que, en algunes ocasions, requereix una conversió més complexa és el tipus C (descriure el comportament d'un ésser viu amb vocabulari propi d'un ens digital), ja que cal explicitar, no només la comparació, sinó com aquesta cal interpretar-la a la inversa del que seria natural interpretar (altres vegades, és suficient en evitar l'ús del terme digital per parlar de propietats humanes).

Proposició original	Proposició final
P59:«A number of people have observed that standard language models don't obviously have a global workspace, but it may be possible to extend them to include a workspace»	P59':«A number of people have observed that standard language models don't obviously have a global workspace **like humans should have if they brain was mechanical theatre**, but it may be possible to extend them to include a workspace»

P140:«exposure to human language in infancy can be thought of as turbocharging the process of acquiring low-level abstractions from sensorimotor interaction»	P141':«exposure to human language in infancy can be thought of as **an acceleration** of the process of acquiring low-level abstractions from sensorimotor interaction»

Per tant, tot i l'especificitat de cada mecanisme, la base conceptual de tots ells, tret d'H i I (que són actitudinals), és una metàfora computacional o comparació que, o bé no és prou explícita o bé simplifica algun dels termes, o bé oblida o amaga part de la informació. En altres paraules, la major part de les estratègies intuïtives per confeccionar una etologia digital es basen en la metàfora computacional o alguna adaptació d'aquesta.

Cal analitzar ara si els mecanismes propis d'una estratègia raonada també tenen com a base conceptual la metàfora computacional. En general, com s'ha observat a l'inici de l'anàlisi, aquesta estratègia sembla molt menys utilitzada, ara bé, cal tenir en compte que les proposicions d'aquest tipus rarament, per si soles, han estat considerades com a vàlides, sinó que s'han agrupat en arguments. Així, per exemple, la P12 no es considera una proposició del mecanisme 10, sinó que es considera que P12-P13-P14-P15 formen part d'un sol argument i les quatre només compten com a una proposició de cara a l'anàlisi quantitatiu. Si no es té en compte aquesta decisió, el nombre total de proposicions d'aquesta estratègia és de 25, entre les quals, el mecanisme més utilitzat és la pretensió de seguir els passos de l'evolució (18 proposicions agrupades en 7 arguments). Majoritàriament, és un mecanisme més utilitzat pels autors de la por (9 proposicions agrupades en 2 arguments per part de Russell i 5 proposicions agrupades en 3 arguments per part de Chalmers, enfront 1 proposició per part de Mitchell i 3 proposicions agrupades en 1 argument per part de Brooks).

El mecanisme de seguir els passos de l'evolució pot ser interpretat de dues maneres, una de dèbil i una de més forta (fent el paral·lelisme amb l'expressió encunyada per Searle). La dèbil és la que utilitza, per exemple, Chalmers en la proposició P61 (modelar el comportament d'un ratolí) o Brooks en P155-P156-P157: ambdós arriben a la conclusió que cal començar a intentar programar allò que, evolutivament, va sorgir primer, com ara la capacitat de moure's per l'entorn sense xocar (el primer Brooks havia intentat programar robots que imitessin el comportament de formigues). Aquests serien els primers passos per arribar a la Lluna, com havia ironitzat Stuart Dreyfus. En canvi, la interpretació forta d'aquest mecanisme –pel qual el projecte IA és literalment comparable a l'evolució, en la mesura que ha d'aconseguir crear uns nous éssers que s'integrin en la naturalesa,

convertint així l'ésser humà en un enginyer capaç de donar vida– despunta en proposicions com P10 de Russell, però cap acadèmic s'atreveix a afirmar-la rotundament, ja que abraçaria les tesis del disseny intel·ligent i contradiria la base fonamental de la teoria de l'evolució. Tanmateix, aquesta interpretació forta és més conseqüent amb la idea implícita en el mecanisme 10: si cal seguir els passos de l'evolució, però s'està dissenyant un artefacte, llavors no té sentit amagar que hom es considera un enginyer o programador que pretén donar vida, és a dir, un Dr. Frankenstein. De moment, no sembla una actitud que ningú, obertament, gosi assumir –sí que, en canvi, s'ha treballat, o com a mínim, s'ha escrit, sobre un tabú similar, el de la immortalitat, ja sigui a partir de traspassar tota la ment al núvol (com defensava Bostrom i acceptava Chalmers), ja sigui també a partir de la cura del cos amb nanorobots (com defensa un dels futuristes que més rèdit ha tret d'explotar la idea de la singularitat, Kurzweil[563]).

Per això, el mecanisme propi de la interpretació forta es podria anomenar la metàfora del programador (la versió informàtica de l'enginyer o del rellotger) en versió antropomòrfica, i la seva forma conceptual no seria molt diferent de la utilitzada en el mecanisme intuïtiu C (descriure el comportament d'un ésser viu amb vocabulari digital). La diferència seria de grau d'explicitació o de conscienciació: mentre que utilitzar el mecanisme C pot ser una forma de parlar, però no necessàriament implica que hom assumeixi la proposició com a literalment certa (pot i sol tenir un grau metafòric), en canvi, la interpretació forta del mecanisme 10 consistiria en assumir que, en la mesura que el programador crea vida (digital), la vida s'ha de comportar com el programador ha decidit crear. Weizenbaum ja va analitzar aquesta eufòria que es pateix quan es programa i un pensa que és capaç de controlar tot allò que crea, especialment si està desenvolupant algun procés de simulació: «The computer programmer's sense of power derives largely from his conviction that his instructions will be obeyed unconditionally and that, given his ability to build arbitrarily large program structures, there is no limit to at least the size of the problems he can solve»[564]. En cert moment, advertia, es pot tenir la temptació de pensar que és el simulador el que va bé i que el problema és que la naturalesa no s'està comportant com toca o estava previst. Precisament aquesta idea és el que aquí s'anomena la inversió de la metàfora computacional o la nova metàfora del programador.

563 KURZWEIL, Ray (13.06.2024). "The Secret to Living Past 120 Years Old? Nanobots" en *Wired*. Consultat el 14 d'agost de 2024 a: https://www.wired.com/story/the-singularity-is-nearer-book-ray-kurzweil/

564 WEIZENBAUM, Joseph (1976). *Computer Power and Human Reason*, Londres, Penguin Books Ltd, New Ed, 1984, pàg. 106.

Aquesta idea, que el programador és amo de la creació, tant dels algoritmes que ha creat com dels que s'ha trobat creats (però que, des d'aquesta perspectiva, no deixen de ser algoritmes), es pot representar com el pas de "l'ordinador és com un cervell" a "el cervell és com un ordinador". A continuació s'analitzaran el requisits conceptuals d'aquesta inversió i, més endavant, les seves conseqüències.

6.2 Els requisits conceptuals d'una etologia digital

Tal i com s'ha vist en el capítol 5, la metàfora computacional és la idea que un ordinador es pot pensar com un cervell, és a dir, que l'ordinador és com un cervell, en aquest ordre. És el segon terme de la comparació, el cervell, el que determina el que es vol predicar del primer terme, l'ordinador. En aquest cas, el cervell s'associa amb l'òrgan del pensament, per tant, quan es diu que l'ordinador és com un cervell el que es vol dir és que l'ordinador pensa.

Aquesta idea se sustenta en la diferència qualitativa entre ambdós termes: el segon terme determina en què consisteix pensar i el primer terme fa l'esforç que faci falta per aconseguir pensar. Això és el que Turing tenia en ment i per això el projecte era aconseguir que un programa d'ordinador imités suficientment bé les respostes d'una persona (pensant) com perquè una altra persona (pensant) acceptés que l'ordinador pensava (aquí, s'obviava que havia estat un programa i no l'ordinador sencer el que havia imitat a la persona). En aquest context, és raonable plantejar-se com dissenyar un programa que imiti com pensen les persones per així poder representar-les millor (és interessant fixar-se aquí també que s'assimili pensar i dir en la mesura que es pretén comprovar que un programa pensa o imita que pensa a través de paraules que s'escriuen a una pantalla, cosa que OpenAI ha entès molt bé: no cal imitar el pensament o ordre de les idees, cal imitar l'ordre de les paraules). Aquest projecte ha tingut tres branques principals: la IA simbòlica, la IA connexionista i la IA encarnada i arrelada.

Sintèticament, tot el projecte de la IA simbòlica, la *Good Old-Fashion AI* (GOFAI), es redueix en traduir les estructures de la lògica a algoritmes. Un dels grans èxits d'aquesta estratègia informàtica han estat els sistemes experts, pels quals, si hom pot pensar i descriure amb claredat un procés, aquest és pot implementar informàticament, ja sigui una partida d'escacs com un sistema de reg que tingui en compte les necessitats de cada planta. Ara bé, des dels anys 60 ja es van plantejar algunes limitacions d'aquest enfocament en vistes a aconseguir que un ordinador simuli pensar, ja que les estructures de la lògica reprodueixen la forma de pensar de la ment, és a dir, el que els humans creiem que fem quan pensem: el vocabulari de la lògica és el vocabulari de la ment, no el vocabulari del cervell. Per això, els primers crítics amb el projecte GOFAI, com Hubert Dreyfus,

van centrar el seu atac en aquesta reducció: una cosa és com ens pensem que pensem i l'altra el que fa el cervell quan pensem.

Aquesta crítica va inaugurar dues tendències: la proposta connexionista, que volia intentar imitar el funcionament del cervell a partir de processos computacionals, i una proposta encarnada i arrelada (*embodied and embedded*), pretesament basada en la fenomenologia de Heidegger o Merleau-Ponty, i que no reduïa la intel·ligència al fet de pensar ni tampoc la situava exclusivament al cervell (per això, més endavant també se l'ha relacionat amb l'enactivisme). El gran èxit de la proposta connexionista han estat els sistemes basats en autoaprenentatge multicapes, com els que permeten l'etiquetatge d'imatges, la conversió de so-paraula-so o imatge-paraula-imatge, els GPTs i, en general, la cerca de patrons en grans quantitats de dades. El gran èxit de l'altra gran proposta, l'encarnada i arrelada, és la robòtica, des de Roomba als gossos robot de Boston Dynamics. La proposta connexionista pretén treballar amb el vocabulari del cervell, i és deutora tant del *big data* i dels èxits de la neurociència com de les seves limitacions. La proposta fenomenològica pretén integrar la visió biològica dels organismes amb la tecnologia sense reduir-la a un intercanvi d'informació (d'alguna forma, la seva inspiració beu més de Wiener que no pas de McCulloch-Shannon).

Ara bé, totes tres branques del projecte IA tenen de base la metàfora computacional i, en algunes ocasions, la seva inversió, és a dir, la metàfora del programador (o de l'enginyer informàtic). Tant la IA simbòlica com la connexionista com l'encarnada neixen sota una determinada manera d'entendre la computació com a informació transformada, és a dir, de la implementació elèctrica de la teoria de la informació. Així, totes tres parteixen de conceptes com *input* i *output* que no apareixen en el text de Turing de 1936, "On computable numbers, with an application to the Entscheidungsproblem", en el qual es descriu per primera vegada una màquina de Turing, base de la mecànica computacional, sinó en el text de Shannon, de 1948, "A Mathematical Theory of Communication". De fet, Turing i Shannon sembla que van coincidir en més d'una ocasió: la primera el 1936, mentre Turing feia el doctorat a Princeton, moment en què va deixar llegir a Shannon el seu article sobre computació; més tard, el 1943, quan Turing va visitar durant dos mesos els laboratoris Bell[565]. Els conceptes de *bit*, *input* i *output* són la conversió de la cel·les de la cinta de Turing a un llenguatge completament discret, ja que tot i que les cel·les eren discretes, la

565 GIANNINI, Tula; BOWEN, Jonathan P. (2017). "Life in Code and Digits: When Shannon met Turing" en *Proceedings of Proceedings of EVA London*. Consultat el 15 d'agost de 2024 a: http://dx.doi.org/10.14236/ewic/EVA2017.9

cinta era contínua i era el que realment es movia, tant cap a l'esquerra com a la dreta (una altra propietat que es perd amb la implementació elèctrica).

Aquesta idea que en la base de la computació hi ha la teoria de la informació, és a dir, que la base comú de les tres branques del projecte IA és el procés d'informació, ha estat el motiu pel qual el psicòleg i fundador del American Institute for Behavioral Research and Technology (AIBRT), Robert Epstein (1953), hagi denominat la metàfora computacional com a metàfora del processament d'informació (*information processing metaphor* o *IP metaphor*)[566]. Aquí, tot i admetent que el requisit conceptual de la metàfora computacional i la seva inversió és la metàfora del processament d'informació, es mantindrà la diferència entre una expressió i l'altra, ja que, precisament, permet tractar aquesta diferència.

El capítol 5 s'ha conclòs amb l'explicitació de les 18 premisses inherents d'aquesta metàfora. A continuació es desglossaran els seus requisits conceptuals i es mostrarà la seva relació amb una etologia digital.

El sentit d'una comparació (premissa 1 i premissa 2)

La premissa 1 i 2 d'una metàfora computacional són propis de qualsevol metàfora, a saber, l'ordre dels elements és rellevant (tot i el que digui Black), i les metàfores tenen vida pròpia, per la qual cosa, poden cosificar-se o enquistar-se, cosa que pot portar a entendre-les de forma literal. Per això Wiener i Rosenblueth havien afirmat que el preu de fer servir una metàfora és haver d'estar vigilant per sempre («The price of metaphor is eternal vigilance»[567]).

Tal i com Konrad Lorenz explicava i s'ha vist en el capítol 1 d'aquest treball, un dels primers problemes que sorgeixen en el camp de l'etologia és el de com descriure el comportament dels animals sense caure en un antropomorfisme. Lorenz ho exemplificava amb la pregunta pel sentit de les ungles del gat i com s'ha de traduir aquesta per ser una pregunta científicament vàlida:

> Cuando preguntamos: "¿para qué el gato tiene uñas puntiagudas curvas y retráctiles?", y contestamos brevemente: "para cazar ratones", esta pregunta no significa en manera alguna que creamos en una predeterminación inherente al universo y a la evolución orgánica. En verdad es

566 EPSTEIN, Robert (18.05.2016). "The empty brain" en *Aeon*. Consultat el 12 d'agost de 2023 a: https://aeon.co/essays/your-brain-does-not-process-information-and-it-is-not-a-computer

567 Citat en LEWONTIN, Richard C. (16.02.2001). "In the Beginning Was the World" en *Science*, Vol. 201, Issue 5507, 2021, pàg. 1264. Consultat el 14 d'agost de 2024 a: https://www.science.org/doi/full/10.1126/science.1057124. No sembla fàcil trobar la citació original de Rosenblueth i Wiener, i alguns que ho han intentat només han aconseguit rastrejar fins un article de Lewontin de 1963 en què ja els hi atribueix la citació.

> una abreviación de la pregunta: "qué función específica es aquella cuyo valor de preservación de la especie confirió a los animales carnívoros (*Felidae*) esa peculiar forma de sus garras?"[568]

Així, l'etologia requereix d'un ús abreviat que permeti agilitzar la comunicació, donant per fet que qualsevol altre lector d'aquest camp sobreentendrà que no s'està intentant denotar cap aspecte teleològic a la pregunta. Aquest ús abreviat (els *shorthands* de Shanahan) és el que permet la metàfora computacional simplement en tant que metàfora: relacionar un sol aspecte del segon terme amb el primer terme. Seria ridícul que algú es prengués massa seriosament, deia Dennett, alguna d'aquestes afirmacions, sense considerar que precisament, prendre-se-les literalment és un dels aspectes inherents a qualsevol metàfora: hi ha qui no entén que les metàfores són metàfores, a l'igual que hi ha qui confon la ficció amb la realitat.

Una etologia digital és possible perquè algú s'ha pres literalment la metàfora computacional.

La metàfora i el model (premissa 3)

La premissa 3 d'una metàfora computacional explicitava el rol del model en la ciència: la relació entre les idees (o teories) i la realitat està mediada per les paraules, en aquest cas, les metàfores. S'ha vist el paper de les metàfores en la ciència i també els perills que comportava, com que, fins i tot si no hi ha una interpretació literal de la metàfora, el vocabulari al qual la metàfora obre la porta incorpori en el camp una visió o una altra, per exemple, la introducció de vocabulari bèl·lic en el camp de la biologia. Així mateix, també s'ha vist com el genocentrisme era conseqüència d'interpretar el codi genètic més com un *program* que no un *programme*, és a dir, de confondre el model per la realitat, que és una forma raonada d'interpretar literalment una metàfora. Aquest raonament s'ha vist explicitat en l'esforç de construir una teleonomia científica diferenciada d'una teleologia precientífica, tant en Mayr com en Wiener.

També s'ha vist que hi ha una diferència entre metàfora i model, perquè mentre la primera està construïda amb un llenguatge natural, els models es solen construir amb un llenguatge formal. Tanmateix, tot i que ambdós representen la realitat, no són la realitat mateixa: «The ultimate model of a cat is of course another cat, whether it be born of still another cat or synthesized in a laboratory»[569] (són inquietants aquestes paraules finals de Wiener si es té en compte que les escriu el 1943). Per evitar aquesta confusió entre representació i realitat, Lewontin defensa que el model ha

568 LORENZ, Konrad (1978). *Fundamentos de la etología. Estudio comparado de las conductas*, Barcelona, Ediciones Paidós, 1986, pàgs. 41-42. La cursiva són en l'original.

569 ROSENBLUETH, Arturo; WIENER, Norbert; BIGELOW, Julian (1943). "Behavior, Purpose and Teleology" en *Philosophy of Science*, Vol. 10, No. 1, gener 1943, pàg. 23.

de ser escollit abans que la metàfora si no es vol condemnar el model a dependre d'un llenguatge figurat: «In sum, what is to be said about the choice of metaphor is that the model should be chosen before the metaphor and not by means of it. The metaphor is to be chosen by virtue of its elements of similarity to the pre-existent structure of rules»[570]. Tanmateix, aquesta exigència no sembla que respecti l'ordre històric dels esdeveniments, si més no en el cas de la metàfora computacional.

La diferència entre una representació formal i una representació natural de la realitat, és a dir, entre el model i la metàfora, i la seva confusió, són en part les responsables de la inversió de la metàfora i, així, part fonamental per construir la versió més perillosa d'una etologia digital, que és la que té efectes negatius sobre la població, no només a curt termini, sinó també a llarg termini, com es veurà més endavant amb l'exemple del *multitasking*.

Pel que fa a la relació de la premissa 3 en la confecció d'una etologia digital, cal valorar-la en funció de la relació d'una etologia digital respecte a una etologia (natural). Lorenz manté la diferència entre model i realitat en especificar constantment que cal observar els animals, tant en llibertat com en captivitat, i que cal conèixer cada espècimen si es vol interpretar correctament el sentit del seu comportament. I aquesta idea preval en la confecció del seu model del Mecanisme Desencadenat Innat, per exemple. Per altra banda, en el camp de l'etologia, un model també pot significar un espècimen fals que serveix per estudiar com es comporta un espècimen natural quan se'l troba. Ara bé, Lorenz, impregnat de la idea teleonòmica, veu una diferència en el camp dels "moderns ordinadors" en el qual un model pot no distingir-se de la realitat, i si ho fa, llavors els seus errors són igualment útils:

> Es precisamente en el terreno de aquellos mecanismos que de la multiplicidad de datos sensoriales extraen las percepciones biológicamente pertinentes, teleonómicas, donde el conocimiento de los modernos ordenadores proporciona más que un mero modelo conceptual en la fisiología del sistema nervioso central.[571]

Per això per a Lorenz es pot fer una excepció amb la computació i, concretament, amb els simuladors, que és a la funció a la qual s'està referint. En la mesura que la teoria de la informació permet introduir un sentit en la biologia, tal i com defensa Mayr explícitament i Lorenz al fer-se seva la proposta teleonòmica, la simulació pot abolir la distància entre model i realitat. Així, una

570 LEWONTIN, Richard C. (1963). "Models, Mathematics and Metaphors" en *Synthese*, Vol. 15, No. 2, juny 1963, pàg. 229. Consultat el 15 d'agost de 2024 a: https://www.jstor.org/stable/20114463

571 LORENZ, Konrad (1978). *Fundamentos de la etología. Estudio comparado de las conductas*, Barcelona, Ediciones Paidós, 1986, pàg. 54.

etologia digital només seria la conseqüència d'interpretar la metàfora computacional des d'una òptica teleonòmica.

La necessitat de dues ignoràncies (premissa 4 – premissa 8)

La premissa 4 reconeix les dues ignoràncies que es requereixen per poder conjugar una metàfora computacional, això és, certa ignorància de com opera el cervell, però també, en el seu moment, de com implementar aquesta ignorància computacionalment, doncs quan es comença a construir l'analogia de l'ordinador amb el cervell, en aquest ordre, tot just s'estan dissenyant els primers ordinadors. Com s'ha dit en la premissa 5 i admetia von Neumann, aquesta ignorància se suplia *prima facie* amb un simplificació: es reduïa el cervell a les seves característiques elèctriques i es decidia ignorar les químiques i les mecàniques. Aquesta ignorància pel que fa al cervell, tot i els avenços en neurociència, encara no s'ha resolt, com testimoniava de Felipe.

La premissa 6 manifesta l'assumpció d'això al mostrar la cura que aquests primers autors tenien en l'expressió de la metàfora i els límits de la seva interpretació, ja sigui utilitzant les cometes o la cursiva per marcar que l'ús d'un terme no es podia prendre en un sentit literal, ja sigui posant en dubte la utilitat mateixa de l'analogia, com per exemple havia fet McCulloch al preguntar-se irònicament si de tant pretendre construir ordinadors imitant el cervell, els ordinadors acabarien patint alguna de les patologies del cervell, com la neurosi. De fet, dos dels comportaments propis d'una etologia digital raonada, la 7 (la pretensió que un programa informàtic pot tenir comportament patològics inesperats) i la 8 (la pretensió que aquests comportaments agafen per sorpresa al seu programador), ja denoten que aquests són uns dels mecanismes que es fan servir en el camp de la informàtica per justificar un error. També Wiener havia constatat els límits del projecte al reconèixer que la informació no és matèria i que la teoria de la informació no convertia els missatges en carn: «The mechanical brain does not secrete thought "as the liver does bile," as the earlier materialists claimed, nor does it put it out in the form of energy, as the muscle puts out its activity. Information is information, not matter or energy. No materialism which does not admit this can survive at the present day».[572]

La premissa 7 justifica el desinterès d'aquests primers autors en la ment: l'analogia és amb el cervell i encara no es diferencia clarament entre el vocabulari de la ment i el vocabulari del cervell. Per una banda es redueix la neurona a un sistema binari i per altra es pretenen aplicar les lleis de la lògica proposicional. La premissa 8 mostra com aquesta síntesi havia estat possible gràcies a

572 WIENER, Norbert (1948). *Cybernetics: Or the Control and Communication in the Animal and the Machine*, Cambridge (MA), MIT Press, 1985, pàg. 144.

l'aplicació de la teoria de la informació com a pont entre el vocabulari de l'ordinador i el vocabulari del cervell, cosa que va transformar les màquines de computar en ordinadors.

El conjunt d'aquestes quatre premisses són el fonament de l'analogia entre l'ordinador i el cervell i, per tant, són arguments necessaris, però no suficients, per construir una etologia digital. De fet, només amb aquestes quatre premisses no seria possible una etologia digital, ja que durant tot aquest període inicial de construcció de l'analogia, l'analogia es manté com a analogia, és a dir, no es produeix la identificació completa entre l'ordinador i el cervell, en altres paraules, la metàfora computacional és llavors encara una comparació. S'està construint un isomorfisme entre el conjunt del vocabulari del cervell i el conjunt del vocabulari de l'ordinador, però sembla que impera la idea que són i seran dos conjunts inherentment diferenciats. Això és el que comença a canviar amb el text de von Neumann que tradueix la premissa 9.

Una concepció matemàtica de la naturalesa (premissa 9)

La premissa 9 expressa que el projecte IA es fonamenta amb cert tipus de realisme matemàtic pel qual l'ordre i connexió de les idees (ordinador) és l'ordre i connexió del món (cervell). Aquesta idea apareixia com a reflexió final del text de von Neumann analitzat en el capítol 5 que tancava la part de construcció de la metàfora computacional i obria el període del seu assentament, si més no, en l'àmbit acadèmic.

Aquest realisme matemàtic no sol ser reconegut com a tal ni per part dels seus partidaris, i se sol presentar en una formulació més dèbil: la matemàtica és el llenguatge que descriu la naturalesa de forma més objectiva. Només alguns autors abracen aquest realisme, com Konrad Zuse al defensar una ontologia digital. La resta s'admiren de la facilitat d'entendre's entre la lògica, la teoria de la informació, la teoria de la computació i el cervell. El que converteix la comparació entre l'ordinador i el cervell en una metàfora, és a dir, el que permet unificar els dos conjunts en un de sol, és aquesta fal·làcia, que denuncia Sabine Hossenfelder en un obra de 2018 titulada *Lost in Math: How Beauty Leads Physics Astray*.

La idea principal contra la que escriu Hossenfelder és, en canvi, un dels requisits argumentals d'una etologia digital, això és, que la matemàtica està tenint / ha de tenir un paper excessivament predominant en altres camps científics fins al punt d'invalidar tesis verificables per falta d'una formulació adequada i validar tesis inverificables gràcies a una bella i elegant formulació: «Y no solo la historia de la ciencia prospera con ideas hermosas que resultaron ser erróneas, sino que, por

otro lado, hay ideas feas que resultaron ser correctas»[573]. Això Hossenfelder ho ataca des del camp de la física, del qual ella n'és especialista en gravetat quàntica, però és també una tendència en qualsevol camp científic i també en el projecte IA.

Hossenfelder centra el seu atac en el següent argument: la bellesa matemàtica no pot ser un criteri de validesa per a una teoria científica (ni tampoc per a una hipòtesi), perquè ni la bellesa és una idea científica ni la matemàtica per si sola pot validar una teoria. Per començar, no hi ha un criteri únic en què consisteix aquesta bellesa, tot i que normalment s'associa a la simplicitat. Ara bé, la simplicitat matemàtica és *ad hoc*: la formulació final és fruit d'un esforç per conjuminar en l'expressió més curta possible, això és, més manejable possible, un conjunt de valors dispersos que es pretén descriure a través d'una equació. El cas de Kepler, que va aconseguir unificar en una sola fórmula suposadament bella i elegant les milers de dades observacionals de les quals disposava, és el paradigma d'aquesta idea. Ara bé, el que demostra aquest cas no és que la naturalesa sigui regular, demostració impossible com va argumentar Hume[574], sinó l'ingeni de Kepler, que va ser capaç d'inventar una fórmula que satisfeia aquell conjunt de dades. I, en la mesura que funcionen (o mentre funcionin), les lleis de Kepler permeten predir la posició d'un planeta amb molta més precisió que propostes anteriors. Per Hossenfelder aquest és el criteri: una teoria científica és vàlida mentre funcioni, més enllà de si algú la troba bella o no: «Todos utilizamos esas matemáticas para computar los resultados de los experimentos y esos cálculos describen correctamente las observaciones. Así es como sabemos que la teoría funciona. De hecho, a eso nos referimos al decir "la teoría funciona"».[575]

La pregunta de si funciona perquè la naturalesa és matemàtica o la matemàtica és una bona eina per descriure la naturalesa, Hossenfelder (i Nima Arkani-Hamed, la persona a la qual entrevista en aquell moment) ho expliquen així:

573 HOSSENFELDER, Sabine (2018). *Perdidos en las matemáticas: cómo la belleza confunde a los físicos*, Barcelona, Editorial Planeta, 2019, pàg. 50.

574 HUME, David (1740). "Abstract of a Book lately Published; Entitled, *A Treatise of Human Nature, &c. Wherein the Chief Argument of that Book is farther Illustrated and Explained"*, pàg. 5. Consultat el 9 d'abril de 2023 a: https://www.earlymoderntexts.com/assets/pdfs/hume1740.pdf. La demostració és impossible perquè com a relació d'idees sempre es pot pensar el contrari (a diferència d'un cercle quadrat), és a dir, no és una idea necessària; i com qüestió de fet, la seva mostració requeriria d'impressions del futur, cosa que és impossible (i suposar que probablement són com les del passat, és circular).

575 HOSSENFELDER, Sabine (2018). *Perdidos en las matemáticas: cómo la belleza confunde a los físicos*, Barcelona, Editorial Planeta, 2019, pàgs. 76-77.

> Usamos esas y otras abstracciones porque funcionan, porque hemos visto que describen la naturaleza. Desde un punto de vista puramente matemático es evidente que no son inevitables; si lo fueran, podríamos deducirlas solo mediante la lógica. Pero nunca podemos demostrar que ningún cálculo matemático sea una verdadera descripción de la naturaleza, ya que las únicas verdades demostrables tienen que ver con las propias estructuras matemáticas, no con la relación de esas estructuras con la realidad.[576]

Aquest també hauria de ser el criteri d'una teoria de la simulació informàtica: una simulació és vàlida si s'ajusta a la realitat, no sí el seu codi és elegant o eficient, que són propietats que mai són sobreres, però no poden determinar la validesa d'una teoria. Ara bé, en el camp de la informàtica hi ha un problema d'indefinició, perquè el mètode i el resultat són el mateix, cosa que pot tergiversar el pes de la prova: com que la matemàtica és el llenguatge de la computació i també el resultat de la simulació, es corre el perill d'oblidar que, fins i tot en el cas que permeti predir el comportament d'una fenomen natural, no són la realitat, sinó un model. I, com explicava Wizenbaum, un model sempre deixa coses a fora:

> What is important in the present context is that models embody only the *essential* features of whatever it is they are intended to represent. […] What aspects of reality are and what are not embodied in a model is entirely a function of the model builder's purpose. But no matter what the purpose, a model, and here I am concerned especially with computer models of aspects of reality, must necessarily *leave out almost everything* that is actually present in the real thing.[577]

La denúncia de Hossenfelder va adreçada al camp de la física teòrica, concretament de la gravetat quàntica, en el qual les verificacions són complexes i depenen d'un conjunt de dades de partícules infinitament petites el qual s'ha hagut de seleccionar prèviament, cosa que pot acabar resultant circular: «Los datos ya no vienen a nosotros; tenemos que saber dónde obtenerlos y no podemos permitirnos buscar en todas partes»[578]. Així, per validar una nova teoria sobre l'univers, cal una hipòtesi prèvia per saber quines dades es faran servir per validar-la, i aquestes dades no només cal que validin les observacions anteriors, sinó que en permetin de noves, cosa que només podrà fer a través d'un experiment que, a part de ser molt car de portar a terme (Hossenfelder té en ment el cost de l'ampliació de l'accelerador de partícules CERN), també pot acabar convertint-se en una

576 *Ibídem*, pàg. 103.

577 WEIZENBAUM, Joseph (1976). *Computer Power and Human Reason*, Londres, Penguin Books Ltd, New Ed, 1984, pàg. XVII. La cursiva és en l'original.

578 HOSSENFELDER, Sabine (2018). *Perdidos en las matemáticas: cómo la belleza confunde a los físicos*, Barcelona, Editorial Planeta, 2019, pàg. 55.

profecia auto-complida: buscar dades que validin una teoria es pot convertir en triar dades que validen la hipòtesi.

Aquesta argumentació circular és en la que cau una etologia digital: com que una simulació és una expressió matemàtica i la matemàtica és el llenguatge de la naturalesa, una simulació matemàtica és, per transitivitat, part de la naturalesa. I si la naturalesa no hi encaixa, el problema és de la naturalesa, no de la simulació, ja que la matemàtica no s'equivoca.

De com la simulació és (premissa 10 - premissa 12)

La premissa 10 precisament expressa aquesta mateixa concepció errònia que no diferencia entre el que Weizenbaum anomena el mode performatiu i el mode teòric, i la base de la qual és una ambigüitat entre formal i objectiu, i entre objectiu i real: com que el llenguatge que s'utilitza per escriure els processos digitals és un llenguatge formal, i una llenguatge formal té les mateixes característiques que les matemàtiques, llavors el resultat d'utilitzar aquest llenguatge és tan objectiu com ho són els resultats matemàtics. Així, la simulació (mode performatiu) és, en la mesura que és objectiva (mode teòric).

Aquí es confon la part pel tot: del fet que el llenguatge utilitzat per descriure o, en aquest cas, simular un fenomen (com pot ser el comportament d'una gota freda o DANA), permeti obtenir un model suficientment similar d'aquest fenomen si més no en una sèrie d'aspectes claus (aspectes que ens faciliten la capacitat de previsió), d'aquí no es pot deduir que aquesta objectivitat es derivi exclusivament del llenguatge utilitzat. De fet, l'objectivitat o no de la simulació derivarà de la fiabilitat amb què s'ha descrit el fenomen en qüestió, ja sigui mitjançant un llenguatge formal ja sigui utilitzant un llenguatge natural. Per exemple, hom pot simular informàticament quelcom molt irreal i, en canvi, descriure objectivament amb llenguatge natural quelcom molt real.

En altres paraules, un procés digital és objectiu si allò que simula s'assembla suficientment a la realitat que pretén representar. En qualsevol cas, la clau de l'objectivitat rau en la simulació i en la representació, dos termes que, precisament, denoten que allò simulat o representat en cap cas pretén ser real, sinó *com si* fos real. Aquest mateix element comparatiu es troba en el concepte "virtual", com Weizenbaum ja va observar: «[…] "it was virtually night", that would have meant, even though it wasn't night it has all the characteristics that a person connects with the night»[579]. El desplaçament semàntic del verb *simular* que s'ha produït fins a convertir l'objecte virtual en un nou

579 WEIZENBAUM, Joseph; WENDT, Gunna (2006). *Islands in the Cyberstream. Seeking Havens of Reason in a Programmed Society*, Litwin Books, Sacramento, CA, 2015, pág. 120.

tipus d'entitat, forma part també d'aquest projecte d'etologia digital, i això ha estat possible per la caiguda d'un "com si". Una realitat virtual és realitat tant com una simulació ho és.

La premissa 11 recull la facilitat de com un usuari final d'un programa es pot confondre a l'utilitzar un programa i acabar associant-li propietats dels éssers vius. Aquesta premissa es deriva de l'anterior: si es pot acceptar que quelcom simulat és, llavors no només un ordinador és com un cervell, sinó que un ordinador és un cervell. La caiguda del "com si" de la premissa 10 es replica en la caiguda del "com" en la premissa 11. Així, tant la secretària de Weizenbaum, que volia intimitat per parlar amb ELIZA, com Blake Lemoine, que defensava que LaMDA era un dolç infant que només volia ajudar, no van tenir una actitud estranya, sinó que la seva actitud era conseqüència d'haver assumit allò simulat com un nou tipus de realitat.

I d'alguna forma, no és d'estranyar, doncs es tracta d'un procés arrelat en els humans, com explicava Marcus referint-s'hi com la bretxa-forat de la credibilitat (*gullibility gap*):

> We attribute intelligence to computers because we have evolved and lived among human beings who themselves base their actions on abstractions like ideas, beliefs, and desires. The behavior of machines is often superficially similar to the behavior of humans, so we are quick to attribute to machines the same sort of underlying mechanisms, even when they lack them.[580]

Marcus associa aquesta transferència d'una teoria de la ment també als artefactes com una derivada del que en psicologia social s'anomena l'error de sobre-atribució fonamental (*fundamental over attribution error*), que consisteix en explicar el comportament de la gent pressuposant un cert tipus de caràcter. Ara bé, també es pot explicar com una correspondència escalar sense fons: en la mesura que es considera que l'*homo sapiens* del davant, que actua o explica quelcom que es reconeix com a propi, també és una persona, i que es reconeixen també en animals algunes actuacions que s'entenen com a pròpies, és comprensible que els infants atribueixin una ment als objectes que reconeixen com a propis. La projecció d'una ment és una característica humana i pot no tenir límits, independentment de l'edat. És un biaix basat en el fet que cada u considera que té una consciència pròpia i que, al mateix temps, no és molt diferent de la persona del costat. Si no es vol caure en un solipsisme, és necessari pressuposar que hi ha altres i que són més o menys similars. Aquesta idea es pot anar aplicant escalonadament fins a adorar una pedra o parlar amb una nina.

Com s'ha vist en Lorenz, l'etologia parteix d'aquesta idea d'escalabilitat entre espècies, cosa que li permet conferir propietats habitualment humanes als animals sense caure en un

580 MARCUS, Gary; DAVIS, Ernest (2019). *Rebooting AI. Building Artificial Intelligence We Can Trust*, Nova York, Vintage Books (Penguin Random House LLC), 2020, pàg. 18.

antropomorfisme. En aquest sentit, l'etologia digital és el mateix, sense el xovinisme de la carn, com l'anomenava Scott Aaronson.[581]

La premissa 12, conseqüència de la premissa 11 i de la premissa 10, conclou que aquesta dinàmica pot i sol portar, no només a una delegació de solucions en la tecnologia, sinó també a una delegació de preguntes i, més preocupant, de responsabilitats. Weizenbaum avisava el 1972 d'aquest possible perill i el 2016 Cathy O'Neil descrivia aquest comportament en l'àmbit judicial, escolar, bancari i laboral, fruit de l'ús d'eines informàtiques que determinaven la probabilitat d'un pres de reincidir, d'un professor de fer millorar els resultats d'una classe, d'un prestatari d'aconseguir un crèdit i, en general, d'una persona d'aconseguir una entrevista de feina. Pel jutge, pel director d'escola, pel prestador o pel contractant, resultava més net i objectiu basar la seva decisió en un càlcul que feia un programa, encara que desconegués les variables utilitzades pel càlcul. O'Neil no només mostrava amb exemples que aquests càlculs estaven sovint esbiaixats, sinó que fins i tot quan no ho estan, tracten la realitat assumint la fiabilitat del model: «If we back away from them and treat mathematical models as a neutral and inevitable force, like the weather or the tides, we abdicate our responsibility. And the result, as we've seen, is WMDs that treat us like machine parts in the workplace, that blackball employees and feast on inequities»[582]. Per *WMD* O'Neil es refereix a *Weapons of Mass Destruction*, l'ús indiscriminat d'eines digitals basades en models contra la població.

Aquest comportament, el de confondre el model per la realitat, ja fou descrit per Hume en el que s'anomena la fal·làcia naturalista i que consisteix en basar l'haver de ser en el ser: en la mesura que s'ha encasellat una persona en cert patró de comportament, fins i tot quan aquest patró no pateix ni parteix de dades i algoritmes esbiaixats (o aquestes han estat estadísticament corregides), aquest patró determina allò que ha de poder fer i, inevitablement, farà aquella persona. Per tant, menysprea la possibilitat del canvi i el dret a canviar, cosa que enquista la societat en un model injust i poc democràtic, que és el que denuncia O'Neil:

> In this march through a virtual lifetime, we've visited school and college, the courts and the workplace, even the voting booth. Along the way, we've witnessed the destruction caused by WMDs. Promising efficiency and fairness, they distort higher education, drive up debt, spur mass incarceration, pummel the poor at nearly every juncture, and undermine democracy. It might seem like the logical response is to disarm these weapons, one by one.[583]

581 AARONSON, Scott (2006). "Lecture 4: Minds and Machines" en *Scott Aaronson*. Consultat el 16 d'agost de 2024 a: https://www.scottaaronson.com/democritus/lec4.html

582 O'NEIL, Cathy (2017). *Armas de destrucción matemática. Cómo el Big Data aumenta la desigualdad y amenaza la democracia*, Capitán Swing Libros, Madrid, 2017, pàg. 247.

583 *Ibídem*, pàg. 269.

En el context d'una etologia digital, aquesta delegació de la responsabilitat no deixa de ser la infantilització que demostra la secretària de Weizenbaum o Blake Lemoine quan acríticament accepten el que una eina o una empresa respectivament els doni per jugar.

La inversió de la metàfora computacional (premissa 13 – premissa 18)

La premissa 13 explica com es comença a produir la inversió de la metàfora computacional per esdevenir el que aquí s'ha anomenat la metàfora del programador (o de l'enginyer informàtic). Aquest inici s'havia analitzat en un text de Dennett en què plantejava quelcom que semblava una contradicció, però no ho era, és a dir, un oxímoron: un camí que no porta enlloc. Aquesta mateixa paradoxa és la base d'una etologia digital: tot i que l'etologia és l'estudi comparatiu del comportament dels éssers vius, és a l'estudiar el comportament d'un ens digital que s'ha programat imitant el comportament dels éssers vius que s'hauria de poder descobrir característiques dels éssers vius que es desconeixen, però que formen part de l'ens digital. Aquesta és la versió més lloable d'una etologia digital i segurament qui l'encarna millor és Gates i la seva il·lusió davant la possibilitat que la IA esdevingui la nova pedra filosofal.

La premissa 14 descriu el procediment d'inversió de la metàfora computacional com un canvi de sentit de la metàfora: l'ordinador ja no és com el cervell sinó que el cervell és com l'ordinador. La conseqüència primera del canvi és que el responsable de que la comparació funcioni és el cervell i l'ordinador és la referència estable: la propietat de pensar (correctament) és de l'ordinador i el cervell és el que fa quelcom de similar. A nivell etològic, aquest canvi es pot explicar de la següent manera: quan s'afirma que un avió (tecnologia) vola com un ocell (naturalesa), s'està comparant la capacitat de sostenir-se a l'aire que s'ha aconseguit donar a l'avió amb la capacitat de sostenir-se a l'aire que té un ocell (i s'obvia o no interessa comparar, per irrellevant, tota la resta d'altres activitats que un ocell pot fer, com ara saltironejar per terra, o capbussar-se a l'aigua, menjar, barallar-se i reproduir-se). El que marca l'element comparatiu és el verb: es compara el fet de volar, és a dir, desplaçar-se per l'aire, i no res més. Estrictament parlant, en la mesura que l'avió es desplaça per l'aire, com fa l'ocell, es pot afirmar, des d'aquest restringit punt de vista, que l'avió vola com un ocell... tot i que és incapaç de fer giravoltes, frenades o diferents piruetes que fan els ocells de veritat. Ara bé, quan s'afirma que és l'ocell el que vola com l'avió, el que fixa la mesura de la comparació és l'avió i el que n'és deutor és l'ocell.

Aquest gir, en principi absurd, que no hauria de tenir més conseqüències, que podria tractar-se només "d'una forma de parlar" que no s'ha de prendre seriosament (aquesta era l'actitud intencional adequada segons Shanahan), acabarà comportant una sèrie de greuges per l'ocell. Si es canvia

l'ocell pels humans i l'avió per l'ordinador, aquestes conseqüències van des de la digitalització d'objectes que no necessiten ser digitalitzats a l'aplicació de teories que provenen del camp de la informàtica al camp de la pedagogia, com el *multitasking*, com s'analitzarà més endavant.

La premissa 15, la premissa 16 i la premissa 17 posen de manifest el conjunt de reduccions que s'han de fer per tal que aquesta inversió funcioni (premissa 15); com això porta que el bit hagi de ser la unitat de mesura comú de la ment i del cervell, és a dir, de les idees i de les coses (premissa 16); i com això es fruit d'una tensió entre l'atzar natural i el sentit humà (premissa 17). Aquestes reduccions, explicades en el capítol 5, són les següents: la intel·ligència es redueix a pensar; pensar es redueix a computar; i computar es redueix a un intercanvi de dades (bits). Un cop el bit és la unitat de mesura de la informació, el projecte IA pateix "els aires de l'època" (no es pot oblidar que és de l'economia d'on prové el vocabulari de la teoria de la informació), que es poden descriure amb dues fórmules, la de l'oportunitat i la de la quantitat.

La fórmula de l'oportunitat la defineix Roberto Casati en una obra titulada *Elogi del paper: contra el colonialisme digital*: «El colonialisme digital és una ideologia que es resumeix en un senzill principi, el condicional: "Es pot, per tant s'ha de". Si és possible que determinada cosa o activitat migri cap al digital, ha de migrar»[584]. La fórmula de la quantitat la defineix Weizenbaum: «if something is good, more is better»[585]. La suma d'ambdues formulacions es tradueix en l'obsessió d'introduir tecnologia en quants més àmbits millor amb la creença que pel sol fet de digitalitzar, ja sigui les aules, els jutjats o els CAPs, els resultats de PISA milloraran, els jutjats no estaran col·lapsats (i es prendran decisions més justes) i els metges tindran temps per mirar el pacient a la cara. La colonització digital és la culminació del projecte modern pel qual totes les qualitats secundàries es poden expressar com a qualitats primàries.

Una etologia digital és un pas més d'aquest projecte general de digitalització, un pas que per alguns, els tecnooptimistes, hauria d'alliberar els humans de les tasques més rutinàries; per altres, pels tecnofòbics, deixarà a l'atur a molta gent. Casati defensa que cal estudiar cas a cas:

> No sóc luddita, sinó anticolonialista. No tot és blanc o negre, sinó que hi ha un ampli ventall de posicions amb infinitat de matisos. Les diferents pràctiques en procés de migració s'hauran de moure, amb més o menys dificultat, entre dos extrems, amb l'acceptació del digital en una banda i la negació a l'altra. Al centre, hi ha un terreny disputat pel qual vaguen el llibre i l'escola, en una perpètua inestabilitat. Es tracta d'un espai que podria revelar-se interessant i ric,

584 CASATI, Roberto (2013). *Elogi del paper. contra el colonialisme digital*, Sant Cugat del Vallès, Pol·len Edicions, 2022, pàg. 14.

585 WEIZENBAUM, Joseph (1976). *Computer Power and Human Reason*, Londres, Penguin Books Ltd, New Ed, 1984, pàg. 27.

> o bé convertir-se en una terra erma, i tot dependrà de com s'acabi negociant la introducció de les tecnologies.[586]

Aquesta posició d'equilibri entre dues tensions, una luddita i l'altra colonialista, una que pretén conservar un món i unes relacions analògiques tal i com estava abans de la revolució industrial i l'altra que pretén canviar-ho tot (segurament per deixar-ho tot igual), curiosament reflecteixen la mateixa tensió que descrivia Mayr entre una teoria de l'evolució atzarosa i la necessitat humana d'un sentit, i beuen de l'habitual dicotomia binària. La síntesi que allà es proposava en forma de teleonomia podria assemblar-se a l'alternativa proposada per Casati[587], més que no a la reacció de Weizenbaum (seria injust, tanmateix, titllar a Weizenbaum de reaccionari, doncs la seva primera reacció, la dels anys 70 i fins la publicació de "Not Without Us" el 1987, va ser pedagògica enlloc de prohibitiva; ara bé, tot i que fins l'última entrevista de 2006 seguia advocant per fer pedagogia, també denunciava sempre que podia que l'ordinador heretava els seus valors dels seus orígens com a eina de guerra[588]).

La premissa 18 reflecteix precisament aquesta possibilitat de l'alternativa transmesa per Casati i Weizenbaum, que en el cas d'una etologia digital ha de tenir en compte diferents factors, el primer dels quals és el que descriu Barry Schwartz:

> Planets don't care what scientists say about their behavior. They move around the sun with complete indifference to how physicists and astronomers theorize about them. Genes are indifferent to our theories about them also. But this is not true of people. Theories about human nature can actually produce changes in how people behave. What this means is that a theory that is false can become true simply by people believing it's true.[589]

Per tant, com que una etologia digital no només és tecnologia, sinó que pretén marcar un patró de comportament, cal considerar també el seu efecte social, pel simple fet que un discurs sobre la

586 CASATI, Roberto (2013). *Elogi del paper. contra el colonialisme digital*, Sant Cugat del Vallès, Pol·len Edicions, 2022, pàg. 134.

587 Una d'aquestes propostes d'equilibri és la d'animar a l'alumnat a escriure a la Viquipèdia, enlloc d'anar-hi només a buscar informació. És una dinàmica que ell utilitzava a les seves classes universitàries, però que el projecte Viquipèdia té pensat també per alumnat de primària i secundària. Quan un alumne aconsegueix que la seva entrada o la seva ampliació o, simplement, una petita correcció (encara que sigui ortogràfica) sigui acceptada per la comunitat, se sent que ha contribuït literalment a la història de la ciència.

588 WEIZENBAUM, Joseph (1992). "Entrevista a Joseph Weizenbaum" en *Telos*, núm.38, Fundación Telefónica. Consultat el 16 d'agost de 2024 a: https://telos.fundaciontelefonica.com/archivo/numero038/entrevista-a-joseph-weizenbaum/

589 SCHWARTZ, Barry (2015). *Why We Work*, TED Books, Simon & Schuster, pàg. 72. Consultat el 16 d'agost de 2024 a: https://archive.org/details/whywework0000schw/mode/

tecnologia també modifica les persones si aquestes creuen que és cert, fins i tot, si l'artefacte en qüestió no ho fa: modifica més el comportament l'amenaça de que ve el llop, que el llop mateix. Aquí s'ha exemplificat amb el discurs de la por i com aquest pretén modificar el comportament, si més no estimular el consum de tecnologia digital. Weizenbaum ho havia exemplificat amb el problema escolar americà. En el següent apartat, s'exemplificarà amb el *multitasking*.

6.3 El *multitasking*

Des de l'àmbit de la psicologia cognitiva, cap a finals dels anys 60 i sobretot els anys 70, es va posar de moda fer servir el concepte de *multitasking* per descriure l'habilitat que té qualsevol humà de fer més d'una cosa aparentment alhora[590]. El concepte havia aparegut anys abans, a mitjans dels anys 60, en el camp de la informàtica quan es van plantejar els primers ordinadors amb més d'un processador: el primer fou l'OS/360 d'IBM, que, posteriorment, va ser considerat també el primer gran fracàs de la indústria del *software*[591]. En el camp de la pedagogia, es va fer servir aquest concepte per estendre la idea que els nadius digitals (un altre concepte desmuntat) eren, no només *multitasking* de forma natural, és a dir, generacionalment, sinó que s'havia d'aprofitar aquest fet per tal d'estimular-los a aprendre més i millor. Ara bé, mentre que el concepte de nadiu digital és popularitzat per Marc Prensky a partir del 2001, el del *multitasking* aplicat a la psicologia i després a la pedagogia el precedeix.

El concepte ha gaudit del seu particular *hype* des de llavors. Mostra d'això poden ser les declaracions d'entre 2001 i 2012 que Michel Desmurget recopila, foteta, per il·lustrar-ho:

> Debemos saber que «en este preciso instante nuestros cerebros están evolucionando a una velocidad nunca antes vista».[19] Además, no nos equivoquemos, nuestros hijos ya no son puramente humanos: se han convertido en «extraterrestres»,[20] en «mutantes»:[20-22] «con una cabeza diferente [...], no vive[n] en el mismo espacio [...], no hablan la misma lengua».[2] «Piensan y procesan la información de un modo esencialmente diferente al de sus predecesores.»[6] «Han nacido con un ratón en una mano y un smartphone en la otra [...], son multitarea, saben hacer de todo y pasan con genialidad de una cosa a otra.»[3] Sus «circuitos neuronales están especialmente cableados para las ciberbúsquedas de fuego rápido».[18] Gracias al efecto beneficioso de todo tipo de pantallas, su cerebro «se desarrolla de un modo diferente».[23] «Ya no [tiene] la misma arquitectura»[24], y en estos momentos está siendo «mejorado,

590 LINDSAY, Peter H.; TAYLOR, Martin M.; FORBES, S.M. (1968). "Attention and multidimensional discrimination" en *Perception & Psychophysics,* volum 4, 1968 pàgs 113–117.

591 CAMPBELL-KELLY, Martin (*et ali.*) (2014). *Computer. A History of the Information Machine*, Routledge, Nova York, 2018, pàgs. 178-192.

aumentado, perfeccionado, amplificado (y liberado) gracias a la tecnología»[25].[592] [les notes en negreta són les de Desmurget]

Per entendre la repercussió de cada una d'aquestes frases, cal tenir en compte que no foren dites per personal poc qualificat en mitjans poc contrastats, sinó persones formades, algunes amb responsabilitats polítiques i, en molts casos, en obres especialitzades o mitjans de comunicació reputats i influents. A continuació es relacionen els noms de les fonts, la professió i l'any de la declaració seguint la numeració de notes de l'original:

[19]: Gary Small, psiquiatre i autor de diferents obres sobre neurociència (2009)

[20]: Jean-Michel Fourgous, exdiputat republicà de l'Assemblea Nacional francesa i autor d'un llibre promovent l'ús d'eines digitals a les escoles (2011)

[21]: Sophie des Déserts, actualment periodista del *Libération* (2012)

[22]: Pascale Nivelle, periodista del *Libération* (2011)

[2]: Michel Serres, filòsof i historiador de la ciència (2011)

[6]: Marc Prensky, escriptor i conferenciant sobre educació (2001)

[3]: Monique Dagnaud, sociòloga i directora emèrita de recerca del Centre Nacional de Recerca Científica (2012)

[18]: novament Gary Small, però en una obra diferent (2011)

[23]: Don Tapscott, empresari i autor d'un llibre sobre creixement personal digital (2009)

[24]: Pierre Kosciusko-Morizet, empresari entrevistat a *Le Figaro* (2012)

[25]: novament Marc Prensky, però en una obra diferent (2012)

Totes elles són representants de la inversió de la metàfora computacional. Aquesta inversió de la metàfora computacional es veu dràsticament aplicada quan és la pedagogia la que s'ha d'adaptar a l'eina, com denuncia Desmurget: «En otras palabras, lo que aquí pongo en tela de juicio son los fundamentos teóricos y las bases experimentales de las desenfrenadas políticas de digitalización del sistema educativo, desde la etapa infantil hasta la universitaria. Lo que cuestiono es esa idea loca de que "la pedagogía debe adaptarse al instrumento [digital]" y no al revés»[593]. Aquesta tendència continua vigent, no només divulgativament entre les famílies, sinó que també en la seva aplicació legislativa.

592 DESMURGET, Michel (2019). *La fábrica de cretinos digitales*, Barcelona, Planeta, 2023, pàg. 17.

593 *Ibídem*, pàg. 234. Desmurget està citant a Emmanuel Davidenkoff, un article titulat "La pédagogie doit s'adapter à l'outil" en *Femme actuelle*, número 1544, 2014. No s'ha pogut trobar aquest article, tot i que sí altres de Davidenkoof en els quals expressa certa preferència per l'ús de la tecnologia a l'aula, sempre i quan sigui supervisat.

Així ho mostra la nova directriu del pla europeu Digital Competence Framework for Educators (DigCompEdu). En la descripció de les competències de l'àrea 3, ensenyament i aprenentatge (*teaching and learning*), la competència 3.1 específica d'ensenyament diu això: «To plan for and implement digital devices and resources in the teaching process, so as to enhance the effectiveness of teaching interventions. To appropriately manage and orchestrate digital teaching interventions. To experiment with and develop new formats and pedagogical methods for instruction»[594]. Per tant, la planificació ja no és en vistes a l'aprenentatge de l'alumnat, sinó a la implementació d'un aparell digital i, encara que l'objectiu sigui la millora de l'efectivitat de les intervencions docents, ningú ha demostrat encara que això s'aconsegueixi precisament a través d'aquesta tecnologia. Encara més preocupant resulta l'expressió "to experiment", precisament en un àmbit en el qual s'hauria d'intentar assegurar l'aprofitament màxim del període educatiu, i deixar els experiments pel laboratori, no pels fills dels altres.

Tanmateix, aquest pla és el que s'està implementat des de fa dos anys a totes escoles i instituts de Catalunya en un projecte denominat Estratègia Digital de Centre, segons el Departament d'Educació, a petició de «Organitzacions internacionals com la UNESCO, l'OCDE i la Comissió Europea, [que] demanen incloure les competències digitals de l'alumnat i del professorat en els objectius dels sistemes educatius (DigComEdu)»[595]. Aquest pla podria ser una oportunitat per millorar l'ús d'aquestes eines entre tota la comunitat educativa, i que tant alumnat com professorat pogués aprendre unes nocions bàsiques d'informàtica així com cert domini de l'ofimàtica, per començar. Tanmateix, aquesta no és la idea final del projecte, que prioritza la innovació pedagògica a la qual creu arribar a través de les eines digitals, el lideratge i l'exhibició *online* dels seus resultats. O això sembla per les respostes a l'enquesta de situació (SELFIE) que cada professor i cada alumne de Catalunya va contestar durant el curs 22-23. A continuació es mostren algunes de les preguntes i les possibles respostes a escollir; aquestes segueixen un ordre des de la més analògica possible (posició 1), a la més digital (posició 7). El resultats totals d'una enquesta omplerta seleccionant la primera opció de cada pregunta, és a dir, la més analògica, es mostren a l'annex 2 (com es pot veure allà, la resposta en la posició número 1 obté 0 punts; en canvi, la resposta en la posició número 7, obté la màxima puntuació).

594 REDECKER, Christine (2017). *European Framework for the Digital Competence of Educators*, Comissió Europea, EUR 28775 EN, pàg. 21. Consultat el 18 d'agost de 2024 a: https://publications.jrc.ec.europa.eu/repository/bitstream/JRC107466/qr/pdf_digcomedu_a4_final.pdf

595 Departament d'Educació de la Generalitat de Catalunya (març 2021). *Estratègia Digital de Centre*, pàg. 5. Consultat el 18 d'agost de 2024 a: https://educacio.gencat.cat/web/.content/home/departament/publicacions/colleccions/pla-educacio-digital/estrategia-digital-centre/estrategia-digital-centre.pdf

* 1 **Valoro con atención cómo, cuándo y por qué usar tecnologías digitales en el aula con mis estudiantes, para garantizar que aporten valor añadido**

- (●) No uso o uso esporádicamente tecnología en mis clases
- ( ) Hago un uso básico del equipamiento disponible (por ejemplo: pizarras digitales, proyectores o entornos de docencia virtual cuando enseño en línea)
- ( ) Utilizo una gran variedad de recursos y herramientas digitales en mis clases
- ( ) Pruebo diferentes métodos de docencia según las tecnologías digitales que elijo
- ( ) Selecciono y pruebo diferentes enfoques de docencia para encontrar los que funcionan mejor para mí
- ( ) Desarrollo mi propio portafolio de actividades, tecnologías y métodos de enseñanza
- ( ) Utilizo herramientas digitales para implementar metodologías docentes innovadoras y compartirlas con mis redes, para que también puedan beneficiarse

Com es pot veure, a partir de la tercera opció, no només es valora que es facin servir «una gran varietat de recursos i eines digitals», sinó que s'adapti els mètodes de docència a les eines, com diu la quarta opció: «Provo diferents mètodes de docència segons les tecnologies digitals que escullo». I aquesta tampoc és l'opció que dóna més punts en l'enquesta, sinó l'última: «Utilitzo eines digitals per implementar metodologies docents innovadores i compartir-les a les meves xarxa, per a què també puguin beneficiar-se».[596]

Una de les altres idees que incorpora l'enquesta és que cal promocionar els treballs en grup i que aquests cal fer-los amb eines digitals.

* 3 **Cuando mis estudiantes trabajan en grupo, utilizan tecnologías digitales para adquirir y plasmar los conocimientos**

- (●) No sé cómo integrar las tecnologías digitales en actividades de aprendizaje colaborativo
- ( ) Integro las tecnologías digitales en actividades de aprendizaje colaborativo
- ( ) Identifico oportunidades e implemento tareas para que los estudiantes trabajen de manera colaborativa buscando información en línea o presentando sus resultados en formatos digitales
- ( ) Estructuro las actividades del curso que requieren que los estudiantes trabajen en colaboración en grupos, utilizando Internet para encontrar información y presentando sus resultados en formatos digitales
- ( ) Diseño tareas de curso que requieren que los estudiantes usen entornos colaborativos en línea para intercambiar conocimiento y debatir
- ( ) Diseño tareas de curso que requieren que los estudiantes usen entornos colaborativos en línea para crear y compartir conocimientos
- ( ) Diseño actividades curriculares que requieren el uso de tecnologías digitales para mejorar el aprendizaje colaborativo y la creación conjunta y el intercambio de conocimientos

Ja no es tracta de «identificar oportunitats i implementar tasques per a què els estudiants treballin de manera col·laborativa buscant informació en línia o presentant els seus resultats en formats digitals» (opció 3), que implicaria prioritzar l'activitat pedagògica i aprofitar per aprendre a fer servir una eina, sinó que el que està més ben valorat per l'enquestador és «dissenyar activitats curriculars que requereixin l'ús de tecnologies digitals per a millorar l'aprenentatge col·laboratiu i la creació conjunta i l'intercanvi de coneixements»[597]. És a dir, el centre de l'aprenentatge no és que ja no sigui el coneixement del docent perquè l'alumne hagi passat a ocupar aquest espai, sinó que

596 "SELFIEforTEACHERS" en *Digital Competence Framework for Educators (DigCompEdu)*. Consultat el 18 d'agost de 2024 a: https://ec.europa.eu/eusurvey/runner/CheckIn_HE_v2021_ES

597 *Ídem.*

tot pivota al voltant de l'eina: tota l'estratègia rau en com maximitzar l'ús de la tecnologia digital, cosa que implica la necessitat d'adquirir-ne més.

Tanmateix, la proposta no només no es basa en evidència científica (com assenyala Gregorio Luri en *La escuela no es un parque de atracciones*, ni els projectes col·laboratius ni els treballs en grup per si sols augmenten els resultats acadèmics[598]), sinó que hi ha literatura acadèmica que la desacredita. Per exemple, des del camp de la neurociència, el *multitaksing* ha quedat completament desacreditat: el cervell només pot bregar amb un fenomen en cada moment i, quan ha de fer diverses coses, simplement en fa primer una i després fa l'altra:

> ¿Realmente somos incapaces de ejecutar dos programas mentales a la vez? A veces tenemos la sensación de que podemos realizar dos tareas distintas o de que el pensamiento puede dividirse y seguir dos líneas distintas en simultáneo, pero esto también es pura ilusión[...]. Cuando los dos objetivos se presentan en simultáneo, la persona realiza la primera tarea a la velocidad normal, pero la segunda se torna sumamente lenta, en proporción directa con el tiempo que insumió tomar la primera decisión (Chun y Marois, 2002; Marti, King y Dehaene, 2015; Marti, Sigman y Dehaene, 2012; Sigman y Dehaene, 2008). En otras palabras, la primera tarea retrasa la segunda: mientras el espacio de trabajo global está ocupado con la primera decisión, la segunda tiene que esperar. Y el retraso es enorme: alcanza fácilmente unos cientos de milisegundos. Si uno está demasiado concentrado en la primera tarea, incluso puede perder por completo el segundo objetivo.[599]

També ha quedat, si no desacreditat, força qüestionat, el benefici de l'ús d'ordinadors i tauletes a les aules en substitució de llibres. Casati, que documenta l'origen, propagadors i fal·làcies de la idea del nadiu digital relacionant-los parcialment amb la proposta d'intel·ligències múltiples de Howard Gardner de l'any 1983, conclou:

> – no existeix una població de "nadius digitals", tret que interpretem el mot "nadius" en un sentit molt superficial i poc interessant;
> – no tenim cap motiu per pensar que existeix una intel·ligència digital específica;
> – per tant, no hem de fer front als suposats problemes d'una població de persones que tindrien una intel·ligència radicalment diferent de la nostra (els extraterrestres no són entre nosaltres);

598 LURI, Gregorio (2022). "¿Existe la «racionalidad pedagógica»?" en *La escuela no es un parque de atracciones*, Barcelona, Ariel, 2022, pàgs. 31-142.

599 DEHAENE, Stanislas (2019). *¿Cómo aprendemos? Los cuatro pilares con los que la educación puede potencias los talentos de nuestro cerebro,* Buenos Aires, SigloXXI Editores, 2019, pàg. 219. Notis l'ús de terminologia provinent del camp de la informàtica fins al punt de ser complicat saber d'on és pròpia, si d'aquest o de la psicologia.

– els efectes positius dels dispositius electrònics sobre el rendiment escolar són molt dubtosos;
– per tant, no hem d'emplenar l'escola de dispositius electrònics per perseguir la il·lusió d'efectes pedagògics en realitat inexistents;
– la multitasca no és una forma d'actuar i pensar, sinó una imposició que es pateix, a causa del mal design i la inèrcia tecnològica, i, per tant,
– s'ha de combatre, no donar-la per feta.[600]

Desmurget també qüestiona l'interès real d'aquestes polítiques davant dels resultats acadèmics aconseguits:

> La segunda tiene que ver con las pantallas en la escuela. También en este caso la literatura científica es clamorosa: cuanto más invierten los países en tecnologías de la información y la comunicación (las célebres TIC) aplicadas a la educación, más baja el rendimiento de los estudiantes. En paralelo, cuanto más tiempo pasan los alumnos con estas tecnologías, más empeoran sus calificaciones. Desde el punto de vista colectivo, estos datos sugieren que el actual movimiento en pro de la digitalización del sistema escolar responde a una lógica más económica que pedagógica.[601]

Desmurget sospita que aquest pla de digitalitzar l'educació té dos objectius: per part de les multinacionals de la tecnologia, vendre; per part dels governs, reduir els costos cada vegada més difícils d'assumir de l'educació. Aquesta teoria, lleugerament conspirativa, podria ser certa si no fos que no només afecta al camp específic de l'educació, ni tampoc només al camp dels serveis públics finançats entre tots, sinó que és una tendència observada en qualsevol àmbit la de cedir una excessiva confiança a una nova eina tecnològica: Gary Marcus i Ernst Davis en *Rebooting IA* expliquen l'excés de confiança que es deposita en els cotxes automàtics[602]; o, com s'ha vist anteriorment, Cathy O'Neill en *Weapons of Math Destruction*, l'excés de confiança d'un jutge quan delega la justificació de la denegació d'una condicional a un *software* especialitzat en recaigudes[603];

600 CASATI, Roberto (2013). *Elogi del paper. contra el colonialisme digital*, Sant Cugat del Vallès, Pol·len Edicions, 2022, pàg. 62. El text original juga amb el canvi de tipologia de lletra en una sèrie de paraules claus. S'ha intentat reproduir marcant aquests termes en Verdana 10.

601 DESMURGET, Michel (2019). *La fábrica de cretinos digitales*, Barcelona, Planeta, 2023, pàg. 254.

602 MARCUS, Gary; DAVIS, Ernest (2019). *Rebooting AI. Building Artificial Intelligence We Can Trust*, Nova York, Vintage books de Penguin Random House, 2020, pàgs. 21-22.

603 O'NEIL, Cathy (2017). "Víctimas civiles: la justicia en la era del *big data*" en *Armas de destrucción matemática. Cómo el Big Data aumenta la desigualdad y amenaza la democracia*, Capitán Swing Libros, Madrid, 2017, pàgs. 107-130.

o Sherry Turckle en *Alone Together*, l'excés de confiança que tothom diposita en les apps per tenir una bona imatge personal i una rica vida social.[604]

Per això, potser no es tracta tant d'una voluntat explícita d'aprofitar-se de la ingenuïtat de ningú, sinó de la ingenuïtat mateixa de tots plegats cap a qualsevol nova tecnologia (no és pot oblidar el Ford Nucleon, que havia de revolucionar l'automobilística a l'utilitzar l'energia nuclear i del qual se'n va arribar a fer un prototip).[605]

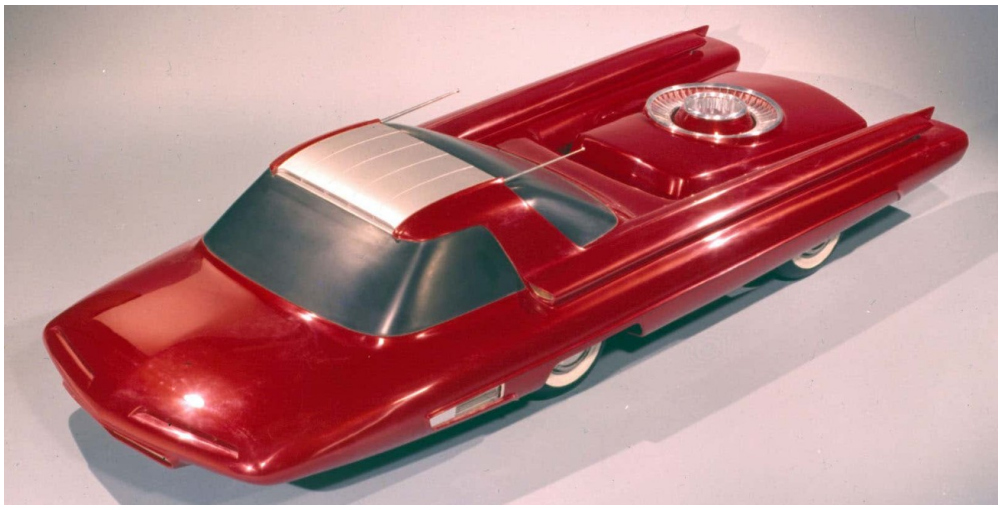

Figura 25: Ford Nucleon, 1958 (Foto: The Drive)

Ingenuïtat, por i il·lusió a parts iguals són els ingredients del *hype* (i segurament algun venedors saben aprofitar-se d'això, fins i tot honradament), especialment en els primers moments; tanmateix, amb el temps, els danys col·laterals són cada vegada més evidents i innegables, i Desmurget posa sobre la taula dades de com està afectant l'ús de pantalles en la capacitat d'atenció dels infants i joves. Ho compara amb casos similars en què, tot i l'evidència científica, per seguir venent un producte (ara ja no tan honradament), es fan mans i mànigues per amagar i qüestionar aquestes evidències, com va passar els anys 60 amb el tabac, els anys 90 amb la pluja àcida[606], o actualment amb l'escalfament global i les begudes ensucrades.[607]

Aquest són els danys col·laterals que, tot i que com a societat ja no s'està disposat a acceptar, perduren fins que els costos de la seva reparació són més alts que els beneficis que generen la seves vendes, i llavors es comença a restringir el consum de tabac i, finalment, s'acaba prohibint.

604 TURKLE, Sherry (2011). *Alone together*, Nova York, Basic Books, 2011. Consultat el 18 d'agost de 2024 a: https://www.mediastudies.asia/wp-content/uploads/2017/02/Sherry_Turkle_Alone_Together.pdf

605 MARQUIS, Erin (17.07.2014). "Nuclear-powered concept cars from the Atomic Age" en *Autoblog*. Consultat el 18 d'agost de 2024 a: https://www.autoblog.com/2014/07/17/nuclear-powered-atomic-age-classic-cars/

606 DESMURGET, Michel (2019). *La fábrica de cretinos digitales*, Barcelona, Planeta, 2023, pàg. 23-24.

607 *Ibídem*, pàg. 64.

Recentment s'està veient l'esforç que suposa la conscienciació global que no s'abusi dels antibiòtics i s'eviti automedicar-se, ja que els costos en sanitat per fer front a la resistència als antimicrobians s'estimen, segons un estudi de la OCDE en 17 països, en 28.9 billons USD[608]. Per tant, el problema real no és el que ve d'aparents canvis immediats, sinó de canvis que es queden durant generacions:

> Sin embargo, hay un tema importante que permanece escondido: que introducimos nuevas cosas en el mundo y éstas tienen un efecto. Hay que decir dos cosas sobre el efecto. Por un lado, que a veces es exponencial. Por otro, que en el mundo, en las innovaciones más profundas, los efectos colaterales llegan a ser más importantes que el efecto original. Pensemos, por ejemplo, en el automóvil.[…] Cuando se introducen ordenadores en la empresa en un primer momento y miras diez años después, el aumento del efecto es bastante pequeño comparado con el tiempo. Pero cuando llegas a otro punto, resulta que en un año el efecto ha sido mucho mayor que en los diez anteriores. Y de repente la situación se hace explosiva.[609]

Tornant a l'exemple de la metàfora computacional, enlloc de preguntar com es pot fer per volar com un ocell (la resposta a la qual fou inventar l'avió), ara es pregunta com es pot fer per aprofitar-se de l'avió (la resposta de la qual ja no és volar com un ocell, sinó explotar les propietats de l'avió que res tenen a veure amb l'ocell, sinó amb els interessos del que l'ha dissenyat). En curt, la inversió de la metàfora computacional és la condició de la possibilitat de l'aprofitament a balquena de la tecnologia digital, independentment dels costos i danys col·laterals que això provoqui, ja sigui l'eliminació d'un estany per allargar una pista d'avió, ja sigui la capacitat d'atenció del jovent de dues generacions. I l'etologia digital és la seva nova estratègia.

608 OCDE (14.09.2023). *Embracing a One Health Framework to Fight Antimicrobial Resistance*, Paris, OECD Health Policy Studies, OECD Publishing, pàg. 13. Consultat el 20 d'agost de 2024 a: https://doi.org/10.1787/ce44c755-en

609 WEIZENBAUM, Joseph (1992). "Entrevista a Joseph Weizenbaum" en *Telos*, núm.38, Fundación Telefónica. Consultat el 16 d'agost de 2024 a: https://telos.fundaciontelefonica.com/archivo/numero038/entrevista-a-joseph-weizenbaum/

Conclusions

La conclusió principal d'aquest treball és que l'etologia digital és una nova estratègia hereva de mecanismes molt més clàssics, com el *hype* i la metàfora. A nivell social té una influència evident a curt termini, amb la proliferació de dispositius digitals i el segrest de l'atenció, tant de grans com de petits. A llarg termini, pot tenir uns costos més alts en educació i model polític. Segurament, com coincideixen Cathy O'Neil i Roberto Casati, cal anar cas a cas i desactivar tots aquells que són perjudicials. Tanmateix, com defensen Michel Desmurget i Joseph Weizenbaum, això potser no és suficient i el que cal és ser més radical i, com a mínim, alçar la veu abans que sigui massa tard.

Una segona conclusió és que tots participem d'una forma o d'una altra en aquesta etologia digital, tant perquè de vegades simplement se'ns amaga sota una forma de parlar tant perquè d'altres és més còmode no discutir-s'hi. Pel que fa a la forma de parlar, s'ha vist que l'antropomorfisme i l'ús de la metàfora són maneres naturals d'expressar-se, i que l'esforç que caldria fer per no caure en aquests usos no sempre podem ni estem disposats a fer-lo: a vegades, també tenim dret a parlar per parlar, és a dir, a no prendre'ns seriosament les nostres paraules. Per altra banda, tampoc es pot estar constantment barallant-se amb el món, i menys quan aquest s'encarna en la figura d'un preadolescent que batalla per no ser l'últim de la classe en tenir mòbil. Però sobretot perquè potser també les noves tecnologies ens generen una estranya fascinació, al mateix temps que ens espanta pensar com canviaran tot el nostre entorn. Vol i dol.

En qualsevol cas, està clar que els discursos simplistes són més fàcils d'entendre que els més raonats, i que quan ens prometen resoldre un problema complex amb una eina miraculosa, tots hi piquem en un moment o altre. Ara bé, com recorda reiteradament Weizenbaum, els problemes humans solen ser problemes socials, i aquests no s'arreglen només amb tecnologia. És cert que un martell pot ajudar a aguantar un quadre, però el problema no és el quadre, sinó la paret en la qual s'ha d'aguantar. Sovint, especialment a l'escola, sembla que les polítiques educatives se centrin més en el martell i el seu color i grandària que no pas en millorar les condicions objectives de l'aula, com la ràtio d'alumnat i les condicions laborals tant de professors com d'estudiants.

A nivell més acadèmic, també s'ha intentat demostrar la relació entre la metàfora computacional i la seva inversió amb la confecció d'una etologia digital. Segurament, una mostra més àmplia de textos i d'autors hagués permès una prova més rodona, però ens sembla que ha quedat prou fonamentada la relació entre aquests discursos i la intenció d'introduir una nova espècie

no animal dins l'arbre taxonòmic[610]. Creiem que cal tenir en compte que el treball no pretenia basar-se en un estudi quantitatiu sinó interpretatiu. La seva representació gràfica, l'objectiu de la qual era facilitar l'explicació, no pot confondre's amb l'estratègia hermenèutica predominant.

També creiem que ha quedat prou clar que tots els membres de la comunitat IA, ja segueixin una estratègia o una altra, tenen per objectius aconseguir replicar la intel·ligència natural en una estructura de silici. I, segurament, tenen més impacte social a llarg termini les paraules d'aquells que ho fan sense caure en discursos catastrofistes, des d'una pràctica més cauta i textos més curosos, que no pas els que opten per traslladar la por a la resta: quan el llop no acaba venint, l'atenció es desvia cap una altra contrada, i torna l'hivern. Potser no és tant perquè com a societat detectem la honestedat i hi confiem, sinó perquè el pas del temps acaba tombant les mentides i els excessos.

Finalment, tot i que la major part de textos i referències puguin ser coneguts pels interessats en aquest camp, s'ha buscat una nova forma de conjuminar-los per presentar una problemàtica general de la filosofia des d'una nova perspectiva. Ens sembla que una de les característiques bàsiques de la filosofia és aquesta voluntat de desmuntar discursos i clixés per intentar entendre com s'han muntat i sobre quins conceptes s'aguanten, no tant per trencar-los, com per posar-los a la vista de tothom, sense que això tingui necessàriament cap altra pretensió que provocar una petita sotragada.

Si aquesta sotragada es produeix i ocasiona un regirament, encara millor, tot i que no és la pretensió d'aquest text.

610 Just acabat aquest treball, va aparèixer un article que es preguntava si la IA tenia una ment, "DOES CHATGPT HAVE A MIND?", de Simon Goldstein i B.A. Levinstein, consultable a: https://arxiv.org/pdf/2407.11015. De fet, ja a finals d'agost de l'any passat es va publicar un altre article, aquest firmat per alguns dels pesos pesants del sector, com Yoshua Bengio i altres, titulat "Consciousness in Artificial Intelligence:Insights from the Science of Consciousness", consultable a: https://arxiv.org/pdf/2308.08708, i que es plantejava la possibilitat que la IA tingués consciència. Sembla que l'interès per seguir utilitzant un discurs etològic passarà per comparar explícitament el nivell de consciència d'alguns animals segons una sèrie d'activitats observables amb les capacitats d'una IA. El fet que David Chalmers hagi estat sempre un defensor de la consciència de certs animals, pot tornar-lo a fer un agradable company de viatge.

7. Taula de figures

Taula de figures

8. Bibliografia

AABY, Anthony A. (1996). “Abstraction and Generalization” en *Internet Archive Wayback Machine*. Consultat, el 18 d’agost de 2023 a: https://web.archive.org/web/20180328151725/http://www.emu.edu.tr:80/aelci/courses/d-318/d-318-files/plbook/absgen.htm

AARONSON, Scott (2006). “Lecture 4: Minds and Machines” en *Scott Aaronson*. Consultat el 16 d’agost de 2024 a: https://www.scottaaronson.com/democritus/lec4.html

ADAM, David (26.04.2024). “Future of Humanity Institute shuts: what’s next for ‘deep future’ research?” en *Nature*. Consultat el 7 de juliol de 2024 a: https://www.nature.com/articles/d41586-024-01229-8

ANDERSON, C.C. (1957). “The latest metaphor in Psychology” en *Dalhousie Review*, Volume 37, Number 2, 1957, pàg. 182. Consultat el 2 d’agost de 2024 a: https://dalspace.library.dal.ca//handle/10222/58774

ANTHONY, Andrew (28.04.2024). “‘Eugenics on steroids’: the toxic and contested legacy of Oxford’s Future of Humanity Institute” en The Guardian. Consultat el 5 de juliol de 2024 a: https://www.theguardian.com/technology/2024/apr/28/nick-bostrom-controversial-future-of-humanity-institute-closure-longtermism-affective-altruism

ARISTÒTIL, *Poètica*, 1457b. Citat en BLACK, Max (1954). “Metaphor” en *Proceedings of the Aristotelian Society*, New Series, Vol. 55 (1954). Consultat el 30 de març de 2024 a: https://www.jstor.org/stable/4544549

BASS, Dina; CLARK, Jack (4.02.2015). "Is Elon Musk Right About AI? Researchers Don't Think So" a *Bloomberg Business*. Consultada el 14 de juliol de 2023 a: https://www.livemint.com/Industry/XAzeyin5n4hI6N98CFI4EJ/The-PR-war-over-artificial-intelligence.html.

BENDER, Emily M.; GEBRU, Timnit (01.03.2021). “On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?” en *FAccT '21: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, març 2021. Consultat el 30 de juliol de 2023 a: https://doi.org/10.1145/3442188.3445922.

BLACK, Max (1954). “Metaphor” en *Proceedings of the Aristotelian Society*, New Series, Vol. 55 (1954). Consultat el 30 de març de 2024 a: https://www.jstor.org/stable/4544549

BLOCK, Ned (1990). “The mind as the software of the brain” en *An Invitation to Cognitive Science: Visual cognition*, Eds. Daniel N. Osherson & Edward E. Smith, Cambridge (MA), MIT Press, 2, 1990, pàgs. 377-425.

BOIX, Xavier; ZEWE, Adam (21.02.2022). “Can machine-learning models overcome biased datasets?” en *MIT News*. Consultat el 15 d’agost de 2023 a: https://news.mit.edu/2022/machine-learning-biased-data-0221

BOYD, Richard N. (1979). “*Metaphor and theory change: What is metaphor" a metaphor for?” en en *Metaphor and thought,* Editor Ortony A, Cambridge, MA, Cambridge University Press, 1979.

BOSTROM, Nick. Nick Bostrom Home’s page. Consultat el 7 de juliol de 2024 a: https://nickbostrom.com/#bio

BOSTROM, Nick. “Curriculum Vitae” en Nick Bostrom Home’s page. Consultat el 8 de juliol de 2024 a: https://nickbostrom.com/cv.pdf

BOSTROM, Nick (2008). “Letter from Utopia” en *Nick Bostrom’s Home Page*. Consultat el 2 d’agost de 2023 a: https://nickbostrom.com/utopia

BOSTROM, Nick (2003). “Are You Living in a Computer Simulation” en *Philosophical Quarterly* (2003), Vol. 53, No. 211, pàg. 244. Consultat el 9 de juliol de 2024 a: https://simulation-argument.com/simulation.pdf

BOSTROM, Nick; SHULMAN, Carl (2023). “Propositions Concerning Digital Minds and Society” properament en *Cambridge Journal of Law, Politics, and Art*, Issue 3, 2024. Consultat el 5 de juliol de 2024 a: https://nickbostrom.com/propositions.pdf

BOYD, Richard N. (1979). “*Metaphor and theory change: What is metaphor" a metaphor for?” en en *Metaphor and thought,* Editor Ortony A, Cambridge, MA, Cambridge University Press, 1979.

BROOKS, Rodney A. (1990). “Elephants Don’t Play Chess” en *Robotics and Autonomous Systems*, 6 (1990). Consultat el 16 de juliol de 2024 a: https://people.csail.mit.edu/brooks/papers/elephants.pdf

BROOKS, Rodney (18.01.2001). “The relationship between matter and life” en *Nature*, vol. 409. Consultat el 17 de juliol de 2024 a: https://www.nature.com/articles/35053196

BROOKS, Rodney A., e*t alii* (13.05.2019). “The Cul-de-Sac of the Computational Metaphor A Talk By Rodney A. Brooks” en *Edge*. Consultat el 30 de juliol de 2023 a: https://www.edge.org/conversation/rodney_a_brooks-the-cul-de-sac-of-the-computational-metaphor

BROOKS, Rodney (23.03.2023). “What Will Transformers Transform?” en *Robots, AI, and other stuff*. Consultat el 5 de juliol de 2024 a: https://rodneybrooks.com/what-will-transformers-transform/

BROOKS, Rodney (08.12.2023). “Three Things That LLMs Have Made Us Rethink” en *Robots, AI, and other stuff.* Consultat el 17 de juliol de 2024 a: https://rodneybrooks.com/three-things-that-llms-have-made-us-rethink/

BROWN, Theodore L. (2008). *Making Truth: Metaphor in Science*, Illinois, University of Illinois Press, 2008, pàgs. 184-185.

CAMPBELL-KELLY, Martin (*et ali.*) (2014). *Computer. A History of the Information Machine*, Routledge, Nova York, 2018.

CELIS, Claudio; SCHULTZ, María Jesús (2021). “Notes on an Algorithmic Faculty of the Imagination” en *Anthropocenes – Human, Inhuman, Posthuman*, 2(1): 12, 2021. Consultat el 17 d’agost de 2023 a: https://www.anthropocenes.net/article/1016/galley/4928/view/.

CERVERÓ MELIÁ, Ernesto; FERRER GISBERT, Pablo; CAPUZ RIZO, Salvador (11-13.07.2018). “The Design Based On Analogies In The Work Of Leonardo Da Vinci” en *22nd International Congress on Project Management and Engineering*. Consultat el 19 de juliol de 2023 a: http://dspace.aeipro.com/xmlui/handle/123456789/1624

CHALMERS, David. “David Chalmers” en *Consc*. Consultat el 10 de juliol de 2024 a: https://consc.net/

CHALMERS, David (2010). “The Singularity: A Philosophical Analysis” en *Journal of Consciousness Studies* 17:7-65, 2010. Consultat el 5 de juliol de 2024 a: https://consc.net/papers/singularity.pdf

CHALMERS, David J. (2023). “Could a Large Language Model be Conscious?” en *arXiv*: 2303.07103. Consultat el 10 de juliol de 2024 a: https://arxiv.org/pdf/2303.07103

CHURCHLAND, Patricia S.; CHURCHLAND, Paul M. (1990). “Could a Machine Think?” en *Scientific American*, 262, 1, 1990, pàg. 37. Consultat el 22 d’agost de 2024 a: http://www.jstor.org/stable/24996642

CORBELLA, Josep; CARBONELL, Eudald; MOYÀ, Salvador; SALA, Robert (2000). *Sapiens: El llarg camí dels homínids cap a la intel·ligència*, Proteus, Barcelona, 2000.

CRAWFORD, Kate (2021). *Atlas of AI. Power, Politic, and the Planetary Costs of Artificial Intelligence*, New Haven, Yale University Press, 2021.

CROSBY , Mathew; BEYRET, Benjamin; SHANAHAN, Murray; HERNÁNDEZ-ORALLO, José; CHEKE, Lucy; HALINA, Marta (2020). “The Animal -AI Testbed an Competition” en *Proceedings of Machine Learning Research*, 123, 2020, pàg. 164.

DANIEL, Will (29.04.2024). “Asana CEO calls Tesla the next Enron and says Elon Musk has misled customers” en *Fortune*. Consultat el 7 de juliol de 2024 a: https://fortune.com/2024/04/29/asana-ceo-tesla-next-enron-elon-musk-misled-customers-investors/

DARWIN, Charles (1859). *On the origin of specie*, http://www.darwin-online.org.uk/. Citat en PIGLIUCCI, Massimo; BOUDRY, Maarten (11.06.2010). “Why Machine-Information Metaphors are Bat for Science and Science Education” en *Sci & Educ* 20, 2011, pàg. 458. Consultat el 23 de juliol de 2024 a: https://doi.org/10.1007/s11191-010-9267-6

“David Chalmers” en *Google Scholar*. Consultat el 10 de juliol de 2024 a: https://scholar.google.com/citations?hl=en&user=o8AfF3MAAAAJ&view_op=list_works

DAVIS, Ernest. *New York University*. Consultat el 14 de juliol de 2024 a: https://cs.nyu.edu/~davise/index.html

DAVIS, Ernest (2021). “Research Papers” en *New York University*. Consultat el 14 de juliol de 2024 a: https://cs.nyu.edu/~davise/pubs.html

DE FELIPE, Javier (2022). *De Laetoli a la Luna: El insólito viaje del cerebro humano*, Barcelona, Crítica, 2022.

DE WAAL, Frans (2016), *¿Tenemos suficiente inteligencia para entender la inteligencia de los animales?*, Barcelona, Tusquets Editores,, 2019.

DENNETT, Daniel C.(1984). “The Role of the Computer Metaphor in Understanding the Mind” en *Annals of the New York Academy of Sciences,* Volume 426, Issue 1, pàg. 267. Consultat el 4 d’agost de 2024 a: https://nyaspubs.onlinelibrary.wiley.com/doi/abs/10.1111/j.1749-6632.1984.tb16524.x

DENNETT, Daniel (1985). “Can machines think?” en *HOW WE KNOW*, editat per MICHAEL SHAFTO a San Francisco, Harper & Row Publishers, 1985, pàgs. 122-145. Consultat el 10 de juliol de 2024 a: https://www.researchgate.net/publication/285475907_Can_Machines_Think

DEHAENE, Stanislas (2019). *¿Cómo aprendemos? Los cuatro pilares con los que la educación puede potencias los talentos de nuestro cerebro,* Buenos Aires, SigloXXI Editores, 2019, pàg. 219.

DENNETT, Danniel (1985). “Can Machines Think” en *How We Know*, Ed. Michael Safto, San Francisco, Harper & Row, 1985, pàg. 140. Consultat el 22 de juliol de 2024 a: https://www.researchgate.net/publication/285475907

Departament d’Educació de la Generalitat de Catalunya (març 2021). *Estratègia Digital de Centre.* Consultat el 18 d’agost de 2024 a: https://educacio.gencat.cat/web/.content/home/departament/publicacions/colleccions/pla-educacio-digital/estrategia-digital-centre/estrategia-digital-centre.pdf

DESMURGET, Michel (2019). *La fábrica de cretinos digitales*, Barcelona, Planeta, 2023.

Diccionario Etimológico Castellano En Línea (DECEL). Consultat el 19 d’agost de 2023 a: https://etimologias.dechile.net/?turborreactor

DREYFUS, Hubert L. (2007). "Why Heideggerian AI Failed and how Fixing it would Require making it more Heideggerian" en *Artificial Intelligence*, 171, 2007. Consultat el 6 de juliol de 2024 a: https://www.sciencedirect.com/science/article/pii/S0004370207001452

DUPRÉ, John (2003). *El legado de Darwin. Qué significa hoy la evolución*, Madrid, Katz Editores, 2009.

ELLUL, Jacques (1954). *La technique ou l'enjeu du siècle*, Paris, 1954 (trad. cast.: *La edad de la técnica*, Barcelona, Ediciones Octaedro, 2003).

EPSTEIN, Robert (18.05.2016). "The empty brain" en *Aeon*. Consultat el 12 d'agost de 2023 a: https://aeon.co/essays/your-brain-does-not-process-information-and-it-is-not-a-computer

FERICEAN , Mihaela Liana; RADA , Olga; BADILITA, Mihaela (2015). "The history and development of ethology" en *Research Journal of Agricultural Science,* 47 (2), 2015.

FLI (2-5.01.2015). "Attendees" en *Future of Life Institute*. Consultat el 15 de juliol de 2023 a: https://futureoflife.org/data/PDF/attendees.pdf

FLI (28.10.2015). "Research Priorities for Robust and Beneficial Artificial Intelligence: An Open Letter" en *Future of Life Institute*. Consultat el 14 de juliol de 2023 a: https://futureoflife.org/open-letter/ai-principles/

FLI (11.08.2017). "AI Principles. The Asilomar AI Principles, coordinated by FLI and developed at the Beneficial AI 2017 conference, are one of the earliest and most influential sets of AI governance principles" en *Future of Life Institute*. Consultat el 13 de juliol de 2023 a: https://futureoflife.org/open-letter/ai-principles/

FLI (14.06.2020)."Foresight in AI Regulation Open Letter" en *Future of Life Institute*. Consultat el 13 de juliol de 2023 a: https://futureoflife.org/open-letter/foresight-in-ai-regulation-open-letter/

FLI (2023). "Our mission" en *Future of Life Institute*. Consultat el 14 de juliol de 2023 a: https://futureoflife.org/our-mission/

FLI (22.03.2023). "Pause Giant AI Experiments: An Open Letter" en *Future of Life Institute*. Consultat el 14 de juliol de 2023 a: https://futureoflife.org/open-letter/pause-giant-ai-experiments/

FLI (31.03.2023). “FAQs about FLI’s Open Letter Calling for a Pause on Giant AI Experiments” en *Future of Life Institute*. Consultat el 13 de juliol de 2023 a: https://futureoflife.org/ai/faqs-about-flis-open-letter-calling-for-a-pause-on-giant-ai-experiments/

FLI (31.03.2023). “Policymaking in the Pause” en *Future of Life Institute*. Consultat el 14 de juliol de 2023 a: https://futureoflife.org/wp-content/uploads/2023/04/FLI_Policymaking_In_The_Pause.pdf

Forbes (27.01.2023). “Microsoft Confirms Its $10 Billion Investment Into ChatGPT, Changing How Microsoft Competes With Google, Apple And Other Tech Giants” en *Forbes*. Consultat el 12 de juliol de 2023 a https://www.forbes.com/sites/qai/2023/01/27/microsoft-confirms-its-10-billion-investment-into-chatgpt-changing-how-microsoft-competes-with-google-apple-and-other-tech-giants/

FORD, Martin (2018). “Rodney Brooks” en *Architects of Intelligence*, Birmingham, Packt Publishing, 2018.

FRANCESCONI, Enrico (2022). “The winter, the summer and the summer dream of artificial intelligence in law: Presidential address to the 18th International Conference on Artificial Intelligence and Law” en *Artificial Intelligence and Law*, 30 (3), DOI: 10.1007/s10506-022-09309-8. Consultat el 16 de juliol de 2024 a: https://doi.org/10.1007/s10506-022-09309-8

“Gary Marcus” en *Substack*. Consultat el 14 de juliol de 2024 a: https://substack.com/@garymarcus

GATES, Bill (21.03.2023). “The Age of AI has begun” en *GatesNotes. The blog of Bill Gates*. Consultat el 6 de juliol de 2023 a: https://www.gatesnotes.com/The-Age-of-AI-Has-Begun.

GEBRU, Timnit; BENDER, Emily M.; MCMILLAN-MAJOR, Angelina; MITCHELL, Margaret (31.03.2023). “Statement from the listed authors of Stochastic Parrots on the «AI pause» letter” en *DAIR Institute*. Consultat el 2 d’agost de 2023 a: https://www.dair-institute.org/blog/letter-statement-March2023/

“George Zarkadakis”. En *Linkedin*. Consultat el 26 d’agost de 2023 a: https://www.linkedin.com/in/gzarkadakis

GIANNINI, Tula; BOWEN, Jonathan P. (2017). “Life in Code and Digits: When Shannon met Turing” en *Proceedings of Proceedings of EVA London*. Consultat el 15 d’agost de 2024 a: http://dx.doi.org/10.14236/ewic/EVA2017.9

GOLDEN, Rebecca (2013). “Mind-Boggling Numbers: Genetic Expression in the Human Brain” en *Science 2.0*, 15.05.2013. Consultat el 8 d’agost de 2022 a: https://www.science20.com/rebecca_goldin/mindboggling_numbers_genetic_expression_human_brain-109345

HALMOS, Paul R. (1973). “The Legend of John von Neumann” en *The American Mathematical Monthly*, Vol. 80, No. 4 (Apr., 1973), pàg. 393. Consultat el 22 de juliol de 2024 a: https://doi.org/10.2307/2319080

HAUGELAND, John (1993). “Mind embodied and embedded” en *Having Thought. Essays in the Metaphysics of Mind,* Cambridge (USA), Harvard University Press, 1998.

HEAVEN, Will Douglas (18.11.2022). “Why Meta’s latest large language model survived only three days online” en *MIT Technological Review*. Consultat el 13 d’agost de 2023 a: https://www.technologyreview.com/2022/11/18/1063487/meta-large-language-model-ai-only-survived-three-days-gpt-3-science/

HOSSENFELDER, Sabine (2018). *Perdidos en las matemáticas: cómo la belleza confunde a los físicos*, Barcelona, Editorial Planeta, 2019.

HUME, David (1740). “Abstract of a Book lately Published; Entitled, *A Treatise of Human Nature, &c. Wherein the Chief Argument of that Book is farther Illustrated and Explained”*. Consultat el 9 d’abril de 2023 a: https://www.earlymoderntexts.com/assets/pdfs/hume1740.pdf

HUME, David (1779). *Dialogues concerning natural religion*, (2nd ed), Hackett. Citat en PIGLIUCCI, Massimo; BOUDRY, Maarten (11.06.2010). “Why Machine-Information Metaphors are Bat for Science and Science Education” en *Sci & Educ* 20, 2011. Consultat el 23 de juliol de 2024 a: https://doi.org/10.1007/s11191-010-9267-6

HUTSON, Matthew (19.11.2019). "'Human Compatible' and 'Artificial Intelligence' Review: Learn Like a Machine" a *The Wall Street Journal*. Consultat el 18 de juliol de 2023 a:

https://www.wsj.com/articles/human-compatible-and-artificial-intelligence-review-learn-like-a-machine-11574207170

Institut d'Investigació en Intel·ligència Artificial del CSIC. Consultat el 14 de juliol de 2023 a: https://www.iiia.csic.es/ca/people/person/?person_id=15

JACOB, François (1973). *The Logic of Life: A History of Heredity*, Nova York, Pantheon Books. Citat en PELUFFO, Alexandre E. (2015). "The "Genetic Program": Behind the Genesis of an Influential Metaphor" en *Genetics,* 200 (3), juliol de 2015, pàg. 688. Consultat el 8 d'agost de 2024 a: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4512536/pdf/685.pdf

KAY, Lily E. (2000). *Who Wrote the Book of Life? A History of the Genetic Code*, Redwood City Stanford University Press. Citat en PELUFFO, Alexandre E. (2015). "The "Genetic Program": Behind the Genesis of an Influential Metaphor" en *Genetics,* 200 (3), juliol de 2015, pàg. 689. Consultat el 8 d'agost de 2024 a: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4512536/pdf/685.pdf

KNOT MAY, Rosalie (25.05.2023). "More than Moore: the next steps for the semiconductor industry" en *Delmic*. Consultat el 27 de juliol de 2024 a: https://blog.delmic.com/more-than-moore-the-next-steps-for-the-semiconductor-industry

KRISHNAMURTHY, Prabhakar (12.09.2019). "Understanding Data Bias" en *Towards Data Science*. Consultat el 15 d'agost de 2023 a: https://towardsdatascience.com/survey-d4f168791e57

KUHN, Thomas S. (1979). "Metaphor in Science" en *Metaphor and thought,* Editor Ortony A, Cambridge, (MA), Cambridge University Press. Consultat el 10 de juliol de 2024 a: https://www.cambridge.org/core/books/abs/metaphor-and-thought/metaphor-in-science/291E8A7ADF9A427260C5C6C8653A1F1F

KURZWEIL, Ray (13.06.2024). "The Secret to Living Past 120 Years Old? Nanobots" en *Wired*. Consultat el 14 d'agost de 2024 a: https://www.wired.com/story/the-singularity-is-nearer-book-ray-kurzweil/

LANIER, Jaron (2010). *You are not a gadget. A manifesto*, Londres, Penguin Books, 2011.

LARSON, Erik J. (2021). *The myth of artificial intelligence: why computers can't think the way we do*, Londres, The Belknap Press of Harvard University Press, 2021.

LESLIE, David (02.10.2019). "Raging robots, hapless humans: the AI dystopia" a *Nature*. Consultat el 18 de juliol de 2023 a: https://www.nature.com/articles/d41586-019-02939-0

LEWONTIN, Richard C. (1963). "Models, Mathematics and Metaphors" en *Synthese*, Vol. 15, No. 2, juny 1963. Consultat el 15 d'agost de 2024 a: https://www.jstor.org/stable/20114463

LEWONTIN, Richard C. (16.02.2001). "In the Beginning Was the World" en *Science*, Vol. 201, Issue 5507, 2021. Consultat el 14 d'agost de 2024 a: https://www.science.org/doi/full/10.1126/science.1057124.

LIEBERMAN, Daniel E. (2010). "Four Legs Good, Two Legs Fortuitous: Brains, Brawn, and the Evolution of Human Bipedalism" en *In the Light of Evolution* (Jonathan B Losos, ed.) Greenwood Village, CO: Roberts & Co, pàg. 16. Consultat el 9 d'agost de 2024 a: https://scholar.harvard.edu/files/dlieberman/files/2010g.pdf

"Lily Kay, 53, life sciences historian" en *MIT News on campus and around the world*. Consultat el 8 d'agost de 2024 a: https://news.mit.edu/2001/kay-0110

LINDSAY, Peter H.; TAYLOR, Martin M.; FORBES, S.M. (1968). "Attention and multidimensional discrimination" en *Perception & Psychophysics,* volum 4, 1968 pàgs 113–117.

LORENZ, Konrad (1978). *Fundamentos de la etología. Estudio comparado de las conductas*, Barcelona, Ediciones Paidós, 1986.

LOVELACE, Augusta Ada (1842). "Nota G" en *On Sketch of the Analytical Engine Invented by Charles Babbage*. Consultat el 17 d'agost de 2023 a: https://notage.org/

LURI, Gregorio (2022). "¿Existe la «racionalidad pedagógica»?" en *La escuela no es un parque de atracciones*, Barcelona, Ariel, 2022, pàgs. 31-142.

LYOTARD, Jean-François (1979). *La condición postmoderna*, Madrid, Cátedra, 2022.

MADDY, Penelope (1993). *Realism in mathematics (A review)*, Londres, Oxford University Press, 1993. Consultat el 27 de juliol de 2024 a: https://web.archive.org/web/20180726044843id_/http://www.ams.org/journals/bull/1995-32-01/S0273-0979-1995-00552-5/S0273-0979-1995-00552-5.pdf

MALL, A. (2003). "analog,digital" en *Theories of Media*, University of Chicago. Consultat el 4 d'agost de 2021: https://csmt.uchicago.edu/glossary2004/analogdigital.htm

MARCUS, Gary; DAVIS, Ernest (2019). *Rebooting AI. Building Artificial Intelligence We Can Trust*, Nova York, Vintage Books (Penguin Random House LLC), 2020.

MARCUS, Gary (2022). *Gary Marcus*. Consultat el 14 de juliol de 2024 a: http://garymarcus.com/

MARCUS, Gary; DAVIS, Ernest (10.01.2023). "Large Language Models like ChatGPT say The Darnedest Things" en *BLOG@CACM*. Consultat el 11 d'agost de 2023 a: https://cacm.acm.org/blogs/blog-cacm/268575-large-language-models-like-chatgpt-say-the-darnedest-things

MARCUS, Gary (08.06.2023). "Two models of AI oversight – and how things could go deeply wrong" en *Marcus on AI*. Consultat el 02 d'agost de 2023 a: https://garymarcus.substack.com/p/two-models-of-ai-oversight-and-how.

MARCUS, Gary; DAVIS, Ernst (18.10.2023). "Reports of the birth of AGI are greatly exaggerated" en *Marcus on AI.* Consultat el 5 de juliol de 2024 a: https://garymarcus.substack.com/p/reports-of-the-birth-of-agi-are-greatly?publication_id=888615&post_id=138077794&isFreemail=true&r=2e3aia

MARCUS, Gary; SOUTHEN, Reid (06.01.2024). "Generative AI Has a Visual Plagiarism Problem" en *IEEE Spectrum*. Consultat el 8 d'agost de 2024 a: https://spectrum.ieee.org/midjourney-copyright

MARCUS, Gary (19.04.2024). "Danniel Dennett, 1942-2024" en *Marcus on AI*. Consultat el 10 de juliol de 2024 a: https://garymarcus.substack.com/p/daniel-dennett-1942-2024

MARCUS, Gary (27.04.2024). “We can’t trust the fox to guard the henhouse, especially when it comes to AI” en *Marcus on AI*. Consultat el 14 de juliol de 2024 a: https://garymarcus.substack.com/p/we-cant-trust-the-fox-to-guard-the

MARCUS, Gary (22.06.2024). “Clarification from Ray Kurzweil” en *Marcus on AI.* Consultat el 9 de juliol de 2024 a: https://garymarcus.substack.com/p/clarification-from-ray-kurzweil

MARQUIS, Erin (17.07.2014). “Nuclear-powered concept cars from the Atomic Age” en *Autoblog*. Consultat el 18 d’agost de 2024 a: https://www.autoblog.com/2014/07/17/nuclear-powered-atomic-age-classic-cars/

MAYR, Ernst (1981). *La biologie de l’évolution.* Citat en PELUFFO, Alexandre E. (2015). “The “Genetic Program”: Behind the Genesis of an Influential Metaphor” en *Genetics,* 200 (3), juliol de 2015, pàg. 687. Consultat el 8 d’agost de 2024 a: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4512536/pdf/685.pdf

MCCARTHY, John; HAYES, Patrick J. (1969). “Some Philosophical Problems from the Standpoint of Artificial Intelligence” en *Machine Intelligence 4,* B. Meltzer & Donald Michie (eds.), Edinburgh University Press., 1969, pàgs. 463-502. Consultat el 18 d’agost de 2023 a: https://www-formal.stanford.edu/jmc/mcchay69.pdf.

MCCULLOCH, Warren S. (1949). “The brain as a computing machine” en *Electrical Engineering*, Volume 68, Issue 6, juny 1949, pàg. 492. Consultat el 2 d’agost de 2024 a: https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=6444817

MCDERMOTT, Drew (abril 1976). “Artificial intelligence meets natural stupidity” en *ACM SIGART Bulletin*, 57, 1976, pàgs. 4–9. Consultat el 21 d’agost de 2024 a: https://doi.org/10.1145/1045339.1045340

MCKAY, Tom (31.07.2017). “No, Facebook Did Not Panic and Shut Down an AI Program That Was Getting Dangerously Smart” en *Gizmodo*. Consultat el 8 d’agost de 2023 a: https://gizmodo.com/no-facebook-did-not-panic-and-shut-down-an-ai-program-1797414922

MCQUILLAN, Dan (21.08.2017). “Data Science as Machinic Neoplatonism” en *Philosophy & Technology*, 31 pàgs. 253–272, 2018. Consultat el 18 d’agost de 2023 a: https://doi.org/10.1007/s13347-017-0273-3

MERCHANT, Brian (31.03.2023). “Afraid of AI? The startups selling it want you to be” en *Los Angeles Times*. Consultat el 2 d’agost de 2023 a: https://www.latimes.com/business/technology/story/2023-03-31/column-afraid-of-ai-the-startups-selling-it-want-you-to-be

MILL, John Stuart (1843). *A System of Logic*. Consultat el 29 de juliol de 2023 a: https://www.laits.utexas.edu/poltheory/mill/sol/sol.b06.c05.html

MINSKY, M. “Decentralized minds” en *Behavioral and Brain Sciences*, 3(3), 1980, pàgs. 439–440. Citat per MITCHELL, Melanie. “Why AI is Harder Than We Think” en *arXiv*. Consultat el 4 d’agost de 2023 a: https://arxiv.org/abs/2104.12871v1

MITCHELL, Melanie (31.10.2019). "We Shouldn't be Scared by 'Superintelligent A.I.'" a *New York Times*. Consultat el 18 de juliol de 2023 a: https://www.nytimes.com/2019/10/31/opinion/superintelligent-artificial-intelligence.html

MITCHELL, Melanie (2019). *Artificial Intelligence. A Guide for Thinking Humans*, Londres, Penguin Random House UK, 2019, pàgs. 67-140.

MITCHELL, Melanie (26.04.2021). “Why AI is Harder Than We Think”, en *arXiv*. Consultat el 30 de juliol de 2023 a: https://arxiv.org/abs/2104.12871

MICTHELL, Melanie; KRAKAUER, David C. (10.02.2023). “The Debate Over Understanding in AI’s Large Language Models” en *arXiv.* Consultat el 22 d’agost de 2023 a: https://arxiv.org/abs/2210.13966

MITCHELL, Melanie (03.04.2023). “Thoughts on a Crazy Week in AI News” en *AI: A Guide for Thinking Humans* (Melanie Mitchell substack). Consultat el 2 d’agost de 2023 a: https://aiguide.substack.com/p/thoughts-on-a-crazy-week-in-ai-news

MORAVEC, Hans (1988). *Mind Children: The Future of Robot and Human Intelligence*, Harvard University Press, 1988, pàgs. 15-16. Citat per MITCHELL, Melanie. “Why AI is Harder Than We Think” en *arXiv*, pàg. 4. Consultat el 4 d’agost de 2023 a: https://arxiv.org/abs/2104.12871v1

MUMFORD, Lewis (1934). *Technics & Civilization*, Nova York, 1934 (trad. Cast.: *Técnica y civilización*, Logroño, Pepitas de calabaza S.L., 2020).

MYERS, Charles S. (1908). *"*Some Observations on the Development of the Colour-sense*"* en *British Journal of Psychology*, Londres, Vol. 2, Iss. 4, (Oct 1, 1908).

NEUMANN, John von (1953). *The Computer & the Brain,* New Haven i Londres, Yale University Press, 2012. Consultat el 22 de juliol de 2024 a: https://ia600707.us.archive.org/3/items/0300181116TheComputerBrain_201901/0300181116_The%20Computer%20Brain.pdf

NILSSON, Nils J. (2010). *The Quest for Artificial Intelligence. A History of Ideas and Achievements*, New York, Cambridge University Press, 2010.

"Nick Bostrom" en *Google Scholar*. Consultat el 8 de juliol de 2024 a: https://scholar.google.com/citations?hl=en&user=oQwpz3QAAAAJ&view_op=list_works

O'NEIL, Cathy (2017). *Armas de destrucción matemática. Cómo el Big Data aumenta la desigualdad y amenaza la democracia*, Madrid, Capitán Swing Libros, 2017.

OCDE (14.09.2023). *Embracing a One Health Framework to Fight Antimicrobial Resistance*, Paris, OECD Health Policy Studies, OECD Publishing. Consultat el 20 d'agost de 2024 a: https://doi.org/10.1787/ce44c755-en

Online Etymology Dictionary. https://www.etymonline.com/word/reward

OpenAi (22.07.2019). "Microsoft invests in and partners with OpenAI to support us building beneficial AGI" en *OpenAi.* Consultat el 12 de juliol a: https://openai.com/blog/microsoft-invests-in-and-partners-with-openai

OpenAI (14.03.2023). "GPT-4" en *OpenAi*. Consultat el 13 d'agost de 2023 a: https://openai.com/research/gpt-4

ORD, Toby (2020). *The Precipice.* Citat en: TORRES, Émile P. (28.07.2021). "The Dangerous Ideas of «Longtermism» and «Existential Risk»" en *Current Affairs*. Consultat el 2 d'agost de 2023 a: https://www.currentaffairs.org/2021/07/the-dangerous-ideas-of-longtermism-and-existential-risk

PASCUAL, Alfredo (29.03.2022). "Singularity se esfuma de España: más de 400 vips, tirados con entradas de 2.500 euros" en *El Confidencial*. Consultat el 10 de juliol de 2024 a:

https://www.elconfidencial.com/empresas/2022-03-29/singularity-university-espana-quiebra-suspension-de-pagos_3396680/

PELÁEZ del Hierro, Fernando; VEÀ BARÓ, Joaquim (1997). *Etología. Bases biológicas de la conducta animal y humana*, Madrid, Ediciones Pirámide, 1997.

PELUFFO, Alexandre E. (2015). "The "Genetic Program": Behind the Genesis of an Influential Metaphor" en *Genetics,* 200 (3), juliol de 2015, pàg. 685. Consultat el 8 d'agost de 2024 a: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4512536/pdf/685.pdf

PINKER, Steven (1997). *How the mind works*, Londres, Penguin, 1998.

PINKER, Steven (10.01.1997). "Organs of Computation: A talk with Steven Pinker" en *Edge*. Consultat el 27 de juliol de 2024 a: https://www.edge.org/conversation/steven_pinker-organs-of-computation

"Rethinking Moral Status" en *Oxford Journals.* Consultat el 9 de juliol de 2024 a: https://oxfordjournals.altmetric.com/details/106356692

RAPAPORT, Willliam J. (1988). "Syntactic Semantics: Foundations of Computational Natural-Language Understanding" en *Aspects of Artificial Intelligence*, James Fetzer (ed.), Kluwer Academic Publishers, Dordrecht, 1988, pàgs. 81-131.

"Research affiliates" en *Center For The Study Of Human Origins*. Consultat el 14 de juliol de 2024 a: https://wp.nyu.edu/csho/people/affiliated_researchers/

ROBINS-EARLY, Nick (20.04.2024). "Oxford shuts down institute run by Elon Musk-backed philosopher" en *The Guardian*. Consultat el 7 de juliol de 2024 a: https://www.theguardian.com/technology/2024/apr/19/oxford-future-of-humanity-institute-closes

"Rodney Brooks – Roboticist" en *MIT Computer Science and Artificial Intelligence Laboratory*. Consultat el 16 de juliol de 2024 a: https://people.csail.mit.edu/brooks/

ROOSE, Kevin (25.02.02023). "El xatbot de Bing va dir que m'estimava i pretenia que deixés la meva dona" originalment en *The New York Times* el 16.02.2023 i traduït pel diari *Ara*. Consultat el 8 de juliol de 2023 a: https://www.ara.cat/economia/tecnologia/xatbot-bing-dir-m-estimava-pretenia-deixes-meva-dona_130_4634725.html

ROOSE, Kevin. *Kevin Rose. Technology writer*. Consultat el 8 de juliol de 2023 a: https://www.kevinroose.com/bio

ROSENBLUETH, Arturo; WIENER,Norbert. Citat en PIGLIUCCI, Massimo; BOUDRY, Maarten (11.06.2010). "Why Machine-Information Metaphors are Bat for Science and Science Education" en *Sci & Educ* 20, 453–471 (2011). Consultat el 23 de juliol de 2024 a: https://doi.org/10.1007/s11191-010-9267-6

ROSENBLUETH, Arturo; WIENER,Norbert; BIGELOW, Julian (1943). "Behavior, Purpose and Teleology" en *Philosophy of Science*, Vol. 10, No. 1, gener, 1943, pàgs. 18-24.

RUSSELL, Stuart; DEWEY, Daniel; TEGMARK, Max (hivern de 2015). "Research Priorities for Robust and Beneficial Artificial Intelligence" en *AI MAGAZINE*. Consultat el 15 de juliol de 2023 a: https://futureoflife.org/data/documents/research_priorities.pdf

RUSSELL, Stuart (02.04.2018). "Specification gaming examples in AI" en *Victoria Krakovna*. Consultat el 22 de juliol de 2023 a: https://vkrakovna.wordpress.com/2018/04/02/specification-gaming-examples-in-ai/.

RUSSELL, Stuart (2019). *Human Compatible. AI and the Problem of Control*, Londres, Penguin Books UK, 2020.

RUSSELL, Stuart (09.03.2021). "Human-Compatible Artificial Intelligence" en *Berkeley Electrical Engineering and Computer Sciences*. Consultat el 26 de juliol de 2023 a: https://people.eecs.berkeley.edu/~russell/papers/mi19book-hcai.pdf

SALIMANS, Tim; HO, Jonathan; CHEN, Xi; SIDOR, Szymon; SUTSKEVER, Ilya (2017). "Evolution strategies as a scalable alternative to reinforcement learning" en *arXiv*, 2017. Consultat el 21 d'agost de 2024 a: arxiv.org/pdf/1703.03864

SAMPLE, Ian (24.10.2019). "Human Compatible by Stuart Russell review – AI and our future" a *The Guardian*. Consultat el 18 de juliol de 2023 a: https://www.theguardian.com/books/2019/oct/24/human-compatible-ai-problem-control-stuart-russell-review

SCHWARTZ, Barry (2015). *Why We Work*, TED Books, Simon & Schuster, pàg. 72. Consultat el 16 d'agost de 2024 a: https://archive.org/details/whywework0000schw/mode/

SEARLE, John R. (1980). “Mind, brains, and programs” en *Behavioral and Brain Sciences*, 3 (3), Cambridge University Press, pàgs. 417-424.

“SELFIEforTEACHERS” en *Digital Competence Framework for Educators (DigCompEdu)*. Consultat el 18 d’agost de 2024 a: https://ec.europa.eu/eusurvey/runner/CheckIn_HE_v2021_ES

SHANAHAN, Murray; MITCHELL, Melanie (29.04.2022). “Abstraction for Deep Reinforcement Learning” en *arXiv*, pàg. 1. Consultat el 4 d’agost de 2023 a: https://arxiv.org/abs/2104.12871v1

SHANAHAN, Murray (16.02.2023). “Talking About Large Language Models” en *arXiv* preprint arXiv:2212.03551. Consultat el 8 d’abril de 2023 a: https://arxiv.org/pdf/2212.03551.pdf

SHANNON, Claude E. (1953). “Computers an Automata” en *Proceedings of the I.R.E.*,Volume: 41, Issue: 10, octubre 1953. Consultat el 3 d’agost de 2024 a: https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=4051186

SHANNON, Claude E. (1992). *Collected Papers*, Ed. N.J.A. Sloane i Aaron D. Wyner, Nova York, IEEE i Wiley Interscience, 1992, pàg. XVII.

SHELLEY, Mary (1818). *Frankenstein; or, The Modern Prometheus*, Cambridge (Massachusetts), The MIT Press, 2017. Consultat el 17 d’agost de 2024 a: https://rauterberg.employee.id.tue.nl/lecturenotes/DDM110%20CAS/Shelley-1818%20Frankenstein.pdf

SHULMAN, Carl; BOSTROM, Nick (2021). “Sharing the World with Digital Minds” en *Rethinking Moral Status,* Clarke, S., Zohny, H. & Savulescu, J. (eds.), Oxford, Oxford University Press, 2021. Consultat el 8 de juliol de 2024 a: https://academic.oup.com/book/41245/chapter/350760172

Stanford Encyclopedia of Philosophy. Consultat el 17 d’agost de 2023 a: https://plato.stanford.edu/entries/frame-problem/

STOUT, Robert L. (1984). “Review of «The role of the computer metaphor in understanding the mind»” en *ACM Digital Library*. Consultat el 5 d’agost de 2024 a: https://dl.acm.org/doi/10.5555/4959.4979

“Stuart Russell”. A *Penguin Random House*. Consultat el 17 de juliol de 2023 a: https://www.penguinrandomhouse.com/authors/2159461/stuart-russell

“Stuart Russell – Biography”, “People @EECS” a *Berkeley Electrical Engineering and Computer Sciences*. Consultat el 18 de juliol de 2023 a: http://people.eecs.berkeley.edu/~russell/biography.html

SYNCED (30.06.2020). “Yann LeCun Quits Twitter Amid Acrimonious Exchanges on AI Bias” en *Synced*. Consultat el 15 d’agost de 2023 a: https://syncedreview.com/2020/06/30/yann-lecun-quits-twitter-amid-acrimonious-exchanges-on-ai-bias/

REDECKER, Christine (2017). *European Framework for the Digital Competence of Educators*, Comissió Europea, EUR 28775 EN. Consultat el 18 d’agost de 2024 a: https://publications.jrc.ec.europa.eu/repository/bitstream/JRC107466/qr/pdf_digcomedu_a4_final.pdf

ROSENBLUETH, Arturo; WIENER, Norbert; BIGELOW, Julian (1943). “Behavior, Purpose and Teleology” en *Philosophy of Science*, Vol. 10, No. 1, gener 1943.

TAYLOR, Cynthia; DEWSBURY, Brryan M. (2018). “On the Problem and Promise of Metaphor Use in Science and Science Communication” en *Journal of Microbiology & Biology Education*, 19(1), 2018. Consultat el 23 de juliol de 2024 a: https://doi.org/10.1128/jmbe.v19i1.1538

TEMPLE-RASTON, Dina (9.10.2015). “The secretive government agency where ‘anything imagined can be tried’” en *The Washington Post*. Consultat el 4 d’agost de 2024 a: https://www.washingtonpost.com/opinions/the-secretive-government-agency-where-anything-imagined-can-be-tried/2015/10/08/3227bc0c-50ce-11e5-933e-7d06c647a395_story.html

TIKU, Nitasha (11.06.2022). “The Google engineer who thinks the company’s AI has come to life” en *The Washington Post*. Consultat el 13 de juny de 2022 a: https://www.washingtonpost.com/technology/2022/06/11/google-ai-lamda-blake-lemoine/

TIMBERLAKE W, Lucas GA. “Behavior system and learning: from misbehavior to general principles”. Citat en MISSLIN, René (27.01.2003). “The defense system of fear: behavior and neurocircuitry” en *Neurophysiologie clinique* 33 (2003), pàgs. 55–66.

“Taming Silicon Valley” en *Penguin Random House*. Consultat el 14 de juliol de 2024 a: https://www.penguinrandomhouse.com/books/768076/taming-silicon-valley-by-gary-f-marcus/

TORRES, Émile P. (28.07.2021). “The Dangerous Ideas of «Longtermism» and «Existential Risk»” en *Current Affairs*. Consultat el 2 d’agost de 2023 a: https://www.currentaffairs.org/2021/07/the-dangerous-ideas-of-longtermism-and-existential-risk

TORRES, Émile P (03.08.2023). “Longtermism poses a real threat to humanity” en *The New Statesman*. Consultat el 8 de juliol de 2024 a: https://www.newstatesman.com/ideas/2023/08/longtermism-threat-humanity

TREFFERT, Darold A. (27.05.2009). “The savant syndrome: an extraordinary condition. A synopsis: past, present, future” en *Philos Trans R Soc Lond B Biol Sci*, 364(1522), 27.05.2009, pàgs. 1351-1357. Consultat el 4 d’agost de 2024 a: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2677584/

TROMP, Nynke; HEKKERT, Paul; VERBEEK, Peter-Paul (2011). “Deign for Socially Responsible Behavior: A classification of Influence Based on Intended User Experience” en *Design Issues.* Volume 27, Number 3, Summer 2011. Consultat el 21 d’agost de 2024 a: https://doi.org/10.1162/DESI_a_00087

TURING, Allan M. (1950). “Computing Machinery and Intelligence”, en *Mind*, núm. 49, pàgs. 433-460 (trad. cast.: ¿Puede pensar una máquina?, Oviedo, KRK Ediciones, 2012).

TURKLE, Sherry (2011). *Alone together*, Nova York, Basic Books, 2011. Consultat el 18 d’agost de 2024 a: https://www.mediastudies.asia/wp-content/uploads/2017/02/Sherry_Turkle_Alone_Together.pdf

ULAM, Stanisław (08.02.1958): “Tribute to John von Neumann” en *Bulletin of the American Mathematical Society*, Vol. 64, No. 3, part 2, 16 maig 1958. pàg. 5. Consultat el 10 de juliol

de 2024 a: https://www.ams.org/journals/bull/1958-64-03/S0002-9904-1958-10189-5/S0002-9904-1958-10189-5.pdf

VINSEL, Lee (01.02.2023). "You're Doing It Wrong: Notes on Criticism and Technology Hype" en *Medium*. Consultat el 27 de juliol de 2023 a: https://sts-news.medium.com/youre-doing-it-wrong-notes-on-criticism-and-technology-hype-18b08b4307e5

WATERS, Richard (18.10.2019). "Human Compatible — can we keep control over a superintelligence?" a *Financial Times*. Consultat el 18 de juliol de 2023 a: https://www.ft.com/content/0e79832c-ef48-11e9-bfa4-b25f11f42901

WEHNER, Mike (31.07.2017). "Facebook engineers panic, pull plug on AI after bots develop their own language" en *BGR*. Consultat el 8 d'agost de 2023 a: https://bgr.com/science/facebook-ai-shutdown-language/

WEIZENBAUM, Joseph (1967). "Contextual Understanding by Computers" en *Communications of the ACM*, Volume 10, Number 8, August 1967, pàgs. 474-478. Consultat el juliol de 2019 a: https://doi.org/10.1145/363534.363545

WEIZENBAUM, Joseph (1972). "On the Impact of the Computer on Society. How does one insult a machine?" en *Science*, vol. 175, maig 1972, pàg. 609. Consultat el 22 d'agost de 2024 a: https://doi.org/10.4000/socio-anthropologie.13631

WEIZENBAUM, Joseph (1976). *Computer Power and Human Reason*, Londres, Penguin Books Ltd, New Ed, 1984.

WEIZENBAUM, Joseph (1992). "Entrevista a Joseph Weizenbaum" en *Telos*, núm.38, Fundación Telefónica. Consultat el 16 d'agost de 2024 a: https://telos.fundaciontelefonica.com/archivo/numero038/entrevista-a-joseph-weizenbaum/

WEIZENBAUM, Joseph; WENDT, Gunna (2006). *Islands in the Cyberstream. Seeking Havens of Reason in a Programmed Society*, Litwin Books, Sacramento, CA, 2015.

Wellshitsguessnot *et alii* (10.01.2024). "Gary Marcus is not an 'AI Expert'" en *Reddit*. Consultat el 16 de juliol de 2024 a: https://www.reddit.com/r/singularity/comments/1931eyy/gary_marcus_is_not_an_ai_expert/

WIENER, Norbert (1948). *Cybernetics: Or the Control and Communication in the Animal and the Machine*, Cambridge (MA), MIT Press, 1985, pàg. 118

ZADOR, Anthony; LECUN, Yann (26.092019). "Don't Fear the Terminator" en *Scientific American*. Consultat el 19 de juliol de 2023 a: https://blogs.scientificamerican.com/observations/dont-fear-the-terminator/

ZARKADAKIS, George (2015). *In our own image*, Londres, Rider Books, 2015.

ZORPETTE, Glenn (17.05.2023). "Just Calm Down About GPT-4 Already And stop confusing performance with competence, says Rodney Brooks" en *IEEE Spectrum*. Consultat el 8 d'agost de 2023 a: https://spectrum.ieee.org/gpt-4-calm-down

9. Agraïments

Aquest treball no hagués estat possible sense el suport i seguiment d'en David Casacuberta, el meu tutor, que ha tingut la gentilesa d'emportar-se de vacances els meus textos i al qual agraeixo que m'hagi fet el retorn fins i tot a mitjans d'agost. Ha estat durant aquest mesos d'estiu que la tesi ha avançat, i això ho he pogut fer perquè podia instal·lar-me a la casa de la muntanya, un dels grans encerts dels meus pares, als que estic infinitament agraït. També a la meva germana, en aquest cas per no molestar massa. I a la Lluïsa i l'Emili, que en moments d'avorriment o capficament, em deixen pujar a desentotsolar-me. Ara bé, algunes de les idees no han sorgit a l'estiu, sinó que són el fruit de moltes hores de classe, intentant explicar algunes coses que segurament jo tampoc entenia del tot, però que havien de tenir sentit, altrament no serien lectures clàssiques. Les converses amb alguns antics alumnes, especialment l'Albert Jiménez, que sempre em dedica els seus treballs (de molta més qualitat i projecció que els meus), el Samir Rezali, el Denis Montoya o l'Andrea Garrido, entre molts d'altres que em sap greu no poder mencionar per manca d'espai, permeten recuperar el gust per aquest ofici. Finalment, al meu fillet, que també està fent el seu treball de recerca, i que segur que és més bonic, i a l'mcrespi, la Mercè, sense les correccions de la qual, no només lingüístiques, no podria ni haver començat.

A tots vosaltres, però també a tots els qui m'he deixat per falta d'espai o de memòria, us saludo amb agraïment.

10. Annexos

10.1 Annex 1: Relació de proposicions

Tot i que segurament sigui innecessari i arbitrari, doncs la interpretació de cada proposició ja s'ha fet en els capítols 2, 3 i 4 d'aquest treball, i tingui poc sentit classificar-les sense tenir en compte el context i la intenció amb la qual foren escrites, a continuació es fa una proposta d'agrupació fruit de l'imperi de la quantificació. S'agrupen les proposicions en tres grups: aquelles que prioritàriament sembla que utilitzen un mecanisme propi de l'estratègia intuïtiva; aquelles que prioritàriament sembla que utilitzen un mecanismes propi de l'estratègia raonada; i aquelles que no s'ha sabut veure que col·laborin en la confecció d'una etologia digital, o no prioritàriament.

Relació de proposicions intuïtives

A continuació es relacionen totes les proposicions que s'han singularitzat en aquest treball segons el mecanisme intuïtiu que fan servir. Algunes proposicions estan en una categoria i és defensable que poden estar en alguna altra; en aquest cas, s'han assignat a la categoria que ha semblat que tenia més pes, però és discutible. Això ha estat possible en tots els casos menys en un, P23, que és tan clar que té elements de dues categories que s'ha optat per fer-la constar en cada una d'elles; per tant, aquesta proposició apareix dues vegades, cosa que fa que el nombre total d'elements relacionats sigui superior al nombre total de proposicions.

A) **Ús antropomòrfic** de termes aplicant-los a un ens digital (i.e., l'ordinador està pensant):

- P2: «Like any rational entity, the algorithm learns how to modify the state of its environment – in this case, the user's mind– in order to maximize its own reward». (Russell)
- P3:«Machines are beneficial to the extent that their actions can be expected to achieve our objectives». (Russell)
- P23:«RL algorithms learn from direct experience of reward signals in the environment, much as a baby learns to stand up from the positive reward of being upright and the negative reward of falling over». (Russell)
- P26:«after seeing Arthur Samuel's checker-playing program learn to play checkers far better than its creator». (Russell)
- P27: «we hope the machine's intelligence will be applied both to learning our true objectives and to helping us achieve them». (Russell)

- P28:«the robot can learn more about human preferences from the observation of human behavior——a process that is the dual of reinforcement learning, wherein behavior is learned from rewards and punishments». (Russell)
- P42:«If an AI is capable of informed consent, then it should not be used to perform work without its informed consent» (Bostrom)
- P45:«The sensory and cognitive capacities of some existing AI systems—and thus their moral status on some accounts—appear in many respects to more closely resemble those of small nonhuman animals than those of typical human adults (on the one hand) or those of rocks or plants (on the other)». (Bostrom)
- P48:«Existing AI is capable of at most quite narrow or rudimentary forms of: abstract and complex thought; self-reflection; deliberation; emotion; creativity and imagination; capacity to think and care about the future in detailed and explicitly temporal ways; long-term and complicated deliberate planning; self-awareness and consciousness of one's own detailed nature; second-order desires; autonomous choice; capacity for deliberative choice; responsiveness to reasons». (Bostrom)
- P89:«AI systems are growing ever more powerful». (Musk4)
- P90:«can learn and adapt after they are sold». (Musk4)
- P96:«AIs also make factual mistakes and experience hallucinations». (Gates)
- P97:«It is learning how to do chat better but can't learn other tasks». (Gates)
- P119:«We need systems that can truly reason about the complex interplay of entities that causally relate to one another in an ever-changing world». (Marcus)
- P132:«Some have questioned why we need machines to have humanlike cognition, but if we want machines to work with us in our human world, we will need them to have the same basic knowledge about the world that is the foundation of our own thinking». (Mitchell)
- P136:«learning statistical associations in the training data that allow the machine to produce correct answers but sometimes for the wrong reasons» (Mitchell)
- P141:«the extent and manner in which machines understand our world has real stakes for how much we can trust them to drive cars, diagnose diseases, care for the elderly, educate children, and more generally act robustly and transparently in tasks that impact humans». (Mitchell)
- P147:«their training in predicting words in vast collections of text has taught them the form of language but not the meaning». (Mitchell)

- P179:«What the large language models are good at is saying what an answer should sound like, which is different from what an answer should be». (Brooks)
- P181:«So we're trying to make our robots human-centered, we call it. They're aware of people. They're using convolutional neural networks to see that that's a person, to see which way they're facing, to see where their legs are, where their arms are». (Brooks)

B) Ús de metàfores o comparacions entre elements analògics i digitals (i.e., l'ordinador funciona com un cervell):

- P5:«Yet every step towards an explanation of how the mind works is also a step towards the creation of the mind's capabilities in an artifact– that is, a step toward artificial intelligence». (Russell)
- P9:«While slow [els cervell humans] compared to electronic circuits, the "cycle time" of a few milliseconds per stage change is fast compared to most biological processes». (Russell)
- P11:«One reason we understand the brain's reward system is that it resembles the method of reinforcement learning developed in AI, for which we have a very solid theory». (Russell)
- P23:«RL algorithms learn from direct experience of reward signals in the environment, much as a baby learns to stand up from the positive reward of being upright and the negative reward of falling over». (Russell)
- P25:«Indeed, one expects to find qualitatively different phenomena occurring when the robot is much less capable than, roughly as capable as, or much more capable than the human». (Russell)
- P46:«One should not fixate too much on "superficial" aspects of an AI system's behavior, appearance, and environment when judging its level of consciousness or moral status: for example, a flexibly intelligent "spreadsheet agent" could share relevant functional and structural properties of a sentient animal even if it lacks a charismatic avatar and is not interacting with natural objects such as food, mates, predators, etc.». (Bostrom)
- P95:«GPT got a 5», «it has aced the test», «it wrote a thoughtful answer», «AI can help». (Gates)
- P114:«The behavior of machines is often superficially similar to the behavior of humans, so we are quick to attribute to machines the same sort of underlying mechanisms, even when they lack them». (Marcus)

- P118:«For real intelligence you also need reasoning, language, and analogy, none of which is nearly so well handled by current technology». (Marcus)
- P120:«The machine-reading systems of our dreams, when they arrive, would be able to answer essentially any reasonable question about what they've read […]. Just as a college student writing a term paper can bring together ideas from multiple sources, cross-validating them and reaching novel conclusions, so too should any machine that can read». (Marcus)
- P121:«Then finally the keystone: construct a kind of human-inspired learning system that uses all the knowledge and cognitive abilities that the AI has; that incorporates what it learns into its prior knowledge; and that, like a child, voraciously learns from every possible source of information: interacting with the world, interacting with people, reading, watching videos, even being explicitly taught. Put all that together, and that's how you get to deep understanding». (Marcus)
- P122:«Computers don't have to work in the same ways as people. There is no need for them to make the many cognitive errors that impair human thought, such as confirmation bias (ignoring data that runs against your prior theories), or to mirror the many limitations of the human mind, such as the difficulty that human beings have in memorizing a list of more than about seven items. There is no reason for machines to do math in the error-prone ways that people do. Humans are flawed in many ways, and machines need not inherit the same limitations. All the same, there is much to be learned from how human minds—which still far outstrip machines when it comes to reading and flexible thinking—work». (Marcus)
- P142:«Moreover, the current debate suggests a fascinating divergence in how to think about understanding in intelligent systems, in particular the contrast between mental models that rely on statistical correlations and those that rely on causal mechanisms». (Mitchell)
- P143:«while AI systems exhibit seemingly intelligent behavior in many specific tasks, they do not understand the data they process in the way humans do». (Micthell)
- P160:«The disciplines of artificial intelligence and artificial life build computational systems inspired by various aspects of life». (Brooks)
- P161:«Researchers in artificial intelligence (AI) and artificial life (Alife) are interested in understanding the properties of living organisms so that they can build artificial systems that exhibit these properties for useful purposes». (Brooks)

- P162:«At the heart of this disappointment lies the fact that neither AI nor Alife has produced artefacts that could be confused with a living organism for more than an instant». (Brooks)

C) **Descriure el comportament d'un ésser viu** amb vocabulari propi d'un ens digital (i.e., els nens són *multitasking*):

- P29: «Indeed, it seems likely that our preferences are at least partially formed by a process resembling inverse reinforcement learning, whereby we absorb preferences that explain the behavior of those around us. Such a process would tend to give cultures some degree of autonomy from the otherwise homogenizing effects of our dopaminebased reward system». (Russell)
- P58:«Second, it's plausible that not all consciousness involves memory, and there may be forms of consciousness which are feedforward» (Chalmers)
- P59:«A number of people have observed that standard language models don't obviously have a global workspace, but it may be possible to extend them to include a workspace» (Chalmers)
- P123:«(As he noted there, each gene is something like an "IF-THEN" statement in a computer program. The THEN side specifies a particular protein to be built, but that protein is only built IF certain chemical signals are available, with each gene having its own unique IF conditions. The result is like an adaptive yet highly compressed set of computer programs, executed autonomously by individual cells, in response to their environments. Learning itself emerges from this stew)». (Marcus)
- P134:«Consider an embodied agent, either an animal (human or nonhuman), a physical robot, or a virtual agent in a simulated environment». (Mitchell)
- P140:«exposure to human language in infancy can be thought of as turbocharging the process of acquiring low-level abstractions from sensorimotor interaction». (Mitchell)

D) **Disminuir capacitats** o atributs d'uns ésser viu per tal que un ens digital s'hi pugui comparar (i.e. els humans tampoc som tan intel·ligents com ens pensem).

- P60:«First: it's arguable that a large degree of disunity is compatible with conscious. Some people are highly disunified, like people with dissociative identity disorders, but they are still conscious. Second: One might argue that a single large language model can support an ecosystem of multiple agents, depending on context, prompting, and the like» (Chalmers)

- P135:«How well an agent does this — that is to say how well it generalises from past experience — is one measure of its intelligence». (Mitchell)
- P137:«We will use the umbrella term abstraction to denote this cluster of operations: seeing similarity (analogy-making), forming a concept, and applying a concept». (Mitchell)
- P138:«Humans and, to a lesser degree, other animals possess a repertoire of these fundamental concepts that includes, in the domain of everyday physics, such concepts as object, path, obstacle, portal, container, and so on, and, in the social domain, such concepts as other agents, being with others, meeting, giving, taking, helping, and so on». (Mitchell)
- P148:«Those on the "LLMs do not understand" side of the debate argue that while the fluency of large language models is surprising, our surprise reflects our lack of intuition of what statistical correlations can produce at the scales of these models». (Mitchell)
- P164:«Building models that are below some complexity threshold also would mean that there is nothing in principle that we do not understand about intelligent or living systems». (Brooks)

E) Desvincular l'eina respecte el seu creador, però això no representa necessàriament una amenaça (i.e. l'aparició de la IA):

- P24:«Bostrom's estimate that superintelligent AI might arrive within this century is roughly consistent with my own, and both are considerably more conservative than those of the typical AI researcher». (Russell)
- P30:«We cannot all be Ruler of the Universe. This means that machines must mediate among conflicting preferences——something that philosophers and social scientists have struggled with for millennia». (Russell)
- P31:«Taking the problem seriously seems likely to yield new ways of thinking about AI, its purpose, and our relationship to it» (Russell)
- P32:«The minds of biological creatures occupy a small corner of a much larger space of possible minds that could be created once we master the technology of artificial intelligence». (Bostrom)
- P39:«Recent rapid advances in artificial intelligence makes it timely to start considering what a future society might look like in which humans share the world with digital minds of various kinds and sophistication». (Bostrom)

- P40:«Some of those digital minds might be sentient or sapient or possess other bases for claiming degrees of moral and/or political status». (Bostrom)
- P41:«At the same time, because their natures may differ in important respects from those of human beings, it would not always be appropriate to simply apply current human norms to such a radically different context». (Bostrom)
- P46:«Some contemporary AI systems (e.g., GPT-3) excel all nonhuman animals in domains such as language, mathematics, and discursive moral argumentation» (Bostrom)
- P49:«So the issue of whether LLMs can be conscious is not the same as the issue of whether they have human-level intelligence. Evolution got to consciousness before it got to human-level consciousness. It's not out of the question that AI might as well». (Chalmers)
- P51:«I'm somewhat skeptical that senses and embodiment are required for consciousness and for understanding. I'd argue that a system with no senses and no body, like the philosopher's classic brain in a vat, could still have conscious thought, even if its consciousness was limited». (Chalmers)
- P71:«how agents that are embedded in their environments should reason». (Musk1)
- P72:«sophisticated agents attempt to manipulate or directly control their reward signals». (Musk1)
- P73:«a system infers the preferences of another rational or nearly rational actor by observing its behavior». (Musk1)
- P74:«another natural subgoal for AI systems pursuing a given goal is the acquisition of fungible resources of a variety of kinds: for example, information about the environment, safety from disruption, and improved freedom of action are all instrumentally useful for many tasks». (Musk1)
- P77:«The emergence of artificial intelligence (AI) promises dramatic changes in our economic and social structures as well as everyday life in Europe and elsewhere; it has been compared to both electricity and the internet». (Musk3)
- P80:«But in each case AI systems of the future will be more capable, more flexible, more general, more continually learning — in short, more intelligent!» (Musk3)
- P86:«and well-resourced institutions for coping with the dramatic economic and political disruptions (especially to democracy) that AI will cause». (Musk4)

- P107:«But none of the breakthroughs of the past few months have moved us substantially closer to strong AI. Artificial intelligence still doesn't control the physical world and can't establish its own goals». (Gates)
- P116:«AI doesn't have to want to destroy us in order to create havoc. In the short term, what we should worry most about is whether machines are actually capable of reliably doing the tasks that we assign them to do». (Marcus)
- P144:«A threshold was reached, as if a space alien suddenly appeared that could communicate with us in an eerily human way. Only one thing is clear—LLMs are not human...Some aspects of their behavior appear to be intelligent, but if not human intelligence, what is the nature of their intelligence?». (Mitchell)
- P146:«The neuroscientist Terrence Sejnowski described the emergence of LLMs this way». (Mitchell)
- P151:«The new methodology bases its decomposition of intelligence into individual behavior generating modules, whose coexistence and co-operation let more complex behaviors emerge». (Brooks)
- P152:«Nouvelle AI relies on the emergence of more global behavior from the interaction of smaller behavioral units». (Brooks)
- P154:«Charmingly, it has been hoped that intelligence will somehow emerge from these simple numeric computations carried out in the sea of symbols». (Brooks)

F) Apel·lar a la inexplicabilitat del seu funcionament (i.e., no sabem com ho fan):

- P50:«But as many people have observed, two decades ago, if we'd seen a system behaving as LLMs do without knowing how it worked, we'd have taken this behavior as fairly strong evidence for intelligence and consciousness» (Chalmers)
- P70:«We recommend expanded research aimed at ensuring that increasingly capable AI systems are robust and beneficial: our AI systems must do what we want them to do» (Musk1)
- P145:«How LLMs perform these feats remains mysterious for lay people and scientists alike». (Mitchell)
- P149:«a long-standing criticism of quantum mechanics is that it provides an effective means of calculation without providing conceptual understanding». (Mitchell)

G) **Assumir que els humans**, de fet, són màquines (i.e., el cervell literalment és un computador):

H) **Presentar la IA com la nova pedra filosofal** que resoldrà els problemes de la humanitat (i.e., la IA resoldrà el canvi climàtic):

- P68:«[...] everything that civilization has to offer is a product of human intelligence; we cannot predict what we might achieve when this intelligence is magnified by the tools AI may provide, but the eradication of disease and poverty are not unfathomable». (Musk1)
- P75:«Artificial intelligence has already provided beneficial tools that are used every day by people around the world. Its continued development, guided by the following principles, will offer amazing opportunities to help and empower people in the decades and centuries ahead» (Musk2)
- P87:«Humanity can enjoy a flourishing future with AI». (Musk4)
- P91:«In my lifetime, I've seen two demonstrations of technology that struck me as revolutionary». (Gates)
- P93:«I've been thinking a lot about how—in addition to helping people be more productive—AI can reduce some of the world's worst inequities». (Gates)
- P98:«[...] these data sets [referint-se a les dades que utilitzen les empreses per formar als seus treballadors] will also be used to train the AIs that will empower people to do this work more efficiently». (Gates)
- P99:«like having a white-collar worker available to help you with various tasks». (Gates)
- P100:«This will both improve your work on the tasks you want to do and free you from the ones you don't want to do». (Gates)
- P101: «Company-wide agents will empower employees in new ways. An agent that understands a particular company will be available for its employees to consult directly and should be part of every meeting so it can answer questions. It can be told to be passive or encouraged to speak up if it has some insight. It will need access to the sales, support, finance, product schedules, and text related to the company. It should read

news related to the industry the company is in. I believe that the result will be that employees will become more productive». (Gates)

- P102:«The rise of AI will free people up to do things that software never will—teaching, caring for patients, and supporting the elderly, for example». (Gates)
- 103:«It will know your interests and your learning style so it can tailor content that will keep you engaged. It will measure your understanding, notice when you're losing interest, and understand what kind of motivation you respond to. It will give immediate feedback». (Gates)
- P104:«Some companies are working on cancer drugs that were developed this way». (Gates)
- P105:«AIs will need a lot of training and further development before they can do things like understand how a certain student learns best or what motivates them». (Gates)
- P106:«New tools will be created for schools that can afford to buy them, but we need to ensure that they are also created for and available to low-income schools in the U.S. and around the world». (Gates)
- P109:«Superintelligent AIs are in our future». (Gates)
- P110:«There will be an explosion of companies working on new uses of AI as well as ways to improve the technology itself». (Gates)
- P111:«There will be immense competition on both approaches». (Gates)
- P112:«No matter what, the subject of AIs will dominate the public discussion for the foreseeable future». (Gates)
- P113:«it's interesting to think about whether artificial intelligence would ever identify inequity and try to reduce it. Do you need to have a sense of morality in order to see inequity, or would a purely rational AI also see it?» (Gates)
- P150:«It could thus be argued that in recent years the field of AI has created machines with new modes of understanding, most likely new species in a larger zoo of related concepts, that will continue to be enriched as we make progress in our pursuit of the elusive nature of intelligence». (Mitchell)

- P153:«Given that neither classical nor nouvelle AI seem close to revealing the secrets of the holy grail of AI, namely general purpose human level intelligence equivalence [...]». (Brooks)
- P158:«The user again is relying on expectations without hard proofs». (Brooks)

I) **Presentar la IA com una nova espècie invasora** amb la qual cal negociar i entendre-s'hi, ja que és una amenaça (i.e., els humans estem en risc d'extinció a causa de la IA):

- P1:«The arrival of superintelligent AI is in many ways analogous to the arrival of a superior alien civilization but much more likely to occur». (Russell)
- P4:«The result will be a new relationship between humans and machines, one that I hope will enable us to navigate the next few decades successfully». (Russell)
- P33:«Here we focus on one set of issues, which arise from the prospect of digital minds with superhumanly strong claims to resources and influence». (Bostrom)
- P34:«Such beings [individual digital minds with superhuman moral status] could contribute immense value to the world, and failing to respect their interests could produce a moral catastrophe, while a naive way of respecting them could be disastrous for humanity». (Bsotrom)
- P35-P36-P37 (Bostrom)
    - P35:«Human biological nature imposes many practical limits on what can be done to promote somebody's welfare. We can only live so long, feel so much joy, have so many children, and benefit so much from additional support and resources».
    - P36:«However, these constraints may loosen for other beings. Consider the possibility of machine minds with conscious experiences, desires, and capacity for reasoning and autonomous decision-making».
    - P37:«This could be a wonderful development: lives free of pain and disease, bubbling over with happiness, enriched with superhuman awareness and understanding and all manner of higher goods»
- P38:«What this means is that, in the long run, total well-being would be much greater to the extent that the world is populated with digital super-beneficiaries rather than life as we know it». (Bostrom)
- P43:«Insofar as future, extraterrestrial, or other civilizations are heavily populated by advanced digital minds, our treatment of the precursors of such minds may be a very important factor in posterity's and ulteriority's assessment of our moral righteousness,

and we have both prudential and moral reasons for taking this perspective into account». (Bostrom)

- P44:«An AI that has high potential to (a) achieve generally superhuman capabilities, and (b) become influential in shaping global outcomes, may have additional claims to moral consideration» (Bostrom)
- P69:«it is important to research how to reap its benefits while avoiding potential pitfalls». (Musk1)
- P76:«Capability Caution: There being no consensus, we should avoid strong assumptions regarding upper limits on future AI capabilities». (Musk2)
- P78:«It is imperative, then, to consider AI not just as it is now, represented largely by a few particular classes of data-driven machine learning systems, but in the forms it is likely to take». (Musk3)
- P79:«AI does and will come in many forms, including as intelligent software tools, as integrated into massive online systems, and as instantiated as software agents designed to substitute for humans». (Musk3)
- P81:«AI systems with human-competitive intelligence can pose profound risks to society and humanity». (Musk4)
- P82:«develop and deploy ever more powerful digital minds that no one – not even their creators – can understand, predict, or reliably control». (Musk4)
- P88:«and give society a chance to adapt». (Musk4)

Relació de proposicions raondades

A continuació es relacionen totes les proposicions que s'han singularitzat en aquest treball segons el mecanisme raonat que fan servir. Al tractar-se de mecanismes raonats, poden aparèixer com un conjunt de proposicions. En aquest cas, s'especifica la sèrie i després es relacionen:

1. La pretensió innecessària de totalitat.

- P85:«a robust auditing and certification ecosystem». (Musk4)
- P133:«In order to understand the nature of true progress in AI, and in particular, why it is harder than we think, we need to move from alchemy to developing a scientific understanding of intelligence». (Mitchell)

2. La pretensió que aquesta totalitat té un sentit o direcció predefinida.

- P124:«In a sort of terminological imperialism, advocates of deep learning often refer to a system, no matter how complex, that contains any deep learning within, as a deep learning system, no matter what role deep learning might play in the larger system, even if other, more traditional elements play a critical role. To us, this seems like calling a car a transmission, just because a transmission plays an important role in the car, or a person a kidney, just because they couldn't live without at least one. Kidneys are obviously critical for human biology, but it doesn't mean that the study of medicine should be reconstrued as nephrology writ large. We anticipate that deep learning will play an important role in hybrid AI systems, but that doesn't mean that they will rely exclusively or even largely on deep learning. Deep learning is much more likely to be a necessary component of intelligence than to be sufficient for intelligence». (Marcus)

3. La pretensió que aquest sentit és natural.

- P17:«One reason artificial intelligence is so fascinating is that it offers a potential route to understanding these issues: we may come to understand both how these intellectual characteristics make intelligent behavior possible and why it's impossible to produce truly intelligent behavior without them». (Russell)
- P125:«Part of the reason we trust other people as much as we do is because we by and large think they will reach the same conclusions as we will, given the same evidence». (Marcus)

4. La pretensió que aquesta naturalitat converteix en una afició innòcua el tracte amb les entitats digitals.

5. La pretensió que un programa ha de ser alliberat per gaudir de la seva màxima expressió.

6. La pretensió que l'actuació d'un programa alliberat permet extreure conclusions que sobrepassen al propi programa, com la de si hi ha trets adquirits o innats.

7. La pretensió que un programa informàtic pot tenir comportaments patològics i inesperats.

- P52-P53-P54 (Chalmers):
  - P52:«On top of this, LLMs have a huge amount of training on text input which derives from sources in the world».
  - P53:«Vision-language models are grounded in images of the environment».

- P54:«In my book on the philosophy of virtual reality, *Reality+*, I've argued that virtual reality is just as legitimate and real as physical reality for all kinds of purposes»

8. La pretensió que els comportaments patològics d'un programa agafen per sorpresa al seu programador.

9. La pretensió que el programador fa una feina similar a una deïtat, ja que crea entitats.

10. La pretensió que el programador ha de seguir els passos de l'evolució per aconseguir que el seu programa pugui inscriure's coherentment a la naturalesa.

- P6-P7-P8-P10 (Russell):
  - P6:«[...] an entity is intelligent to the extent that what it does is likely to achieve what it wants, given what it has perceived».
  - P7:«Evolution doesn't know, in advance, where the glucose is going to be or where your keys are, so putting the capability to find them into the organism is the next best thing».
  - P8:«It [*E. coli*] never learns. It has no brain, just a few simple chemical reactions to do the job».
  - P10:«the neural implementation of the *cognitive* level –learning, knowing, remembering, reasoning, planning, deciding, and so on– is still mostly anyone's guest. (Perhaps that will change as we understand more about AI [...])».
- P12-P13-P14-P15-P16 (Russell):
  - P12:«Learning is good for more than surviving and prospering. It also *speeds up evolution*».
  - P13:«After all, learning doesn't change one's DNA, and evolution is all about changing DNA over generations».
  - P14:«Clearly, if evolution has to worry about choosing only the first three digits, its job is much easier; the adaptive organism, in learning the last three digits, is doing in one lifetime what evolution would have taken many generations to do».
  - P15:«"How did the reward system get there in the first place?" The answer, of course, is by an evolutionary process, one that internalized a feedback mechanism that is at least somewhat aligned with evolutionary fitness».
  - P16:«Evolution considers you only as an agent, that is, something that acts».

- P55-P56-P57 (Chalmers):
  - P55:«One key idea here is that world-models are just modeling text and not modeling the world. They don't have genuine understanding and meaning of the kind you get from a genuine world-model».
  - P56:«I think it's important to make a distinction between training and (post-training) online processing here. It's true that LLMs are trained to minimize prediction error in string matching, but that doesn't mean that their processing is just string matching».
  - P57:«An analogy: in evolution by natural selection, maximizing fitness during evolution can lead to wholly novel processes post-evolution. A critic might say, all these systems are doing is maximizing fitness. But it turns out that the best way for organisms to maximize fitness is to have these amazing capacities – like seeing and flying and even having world-models. Likewise, it may well turn out that the best way for a system to minimize prediction error during training is for it to use highly novel processes, including world-models»
- P61:«But one can also try to model the perception-action cycle of, say, a single mouse». (Chalmers)
- P62:«If we reach that point, there would be a serious chance that those systems are conscious. Multiplying those chances gives us a significant chance of at least mouse-level consciousness with a decade». (Chalmers)
- P139:«several years' experience observing and interacting with the real world in all its variety and complexity, not to mention the innate endowment of human evolutionary history» (Mitchell)
- P155-P156-P157 (Brooks):
  - P155:«We already have an existence proof of the possibility of intelligent entities — human beings. Additionally many animals are intelligent to some degree (...). They have evolved over the 4.6 billion year history of the earth […]».
  - P156:«It is instructive to reflect on the way in which earth-based biological evolution spent its time».
  - P157:«That essence is the ability to move around in a dynamic environment, sensing the surroundings to a degree sufficient to achieve the necessary maintenance of life and reproduction. This part of intelligence is where evolution has concentrated its time—it is much harder. This is the physically grounded part of animal systems».

Proposicions sense connotació etològica

A continuació es relacionen les proposicions que, aparentment, no tenen cap paper en la confecció d'una etologia. Tanmateix, algunes seria discutible no tant per la proposició mateixa com per la intenció del context en el qual foren formulades. Ara bé, com que aquest context no era suficientment evident, s'ha optat per incloure-les en aquest apartat juntament amb altres que clarament defensen un paper no etològic de la tecnologia.

- P18:«Although comparisons between computers and brains are not especially meaningful, the numbers for Summit slightly exceed the raw capacity of the human brain[...]». (Russell)
- P19:«Even in the 1950s, computers were described in the popular press as "super-brains" that were "faster than Einstein." So can we say now, finally, that computers are as powerful as the human brain? No. Focusing on raw computing power misses the point entirely». (Russell)
- P20:«Contrary to common interpretations, I doubt that the test was intended as a true definition of intelligence, in the sense that a machine is intelligent if and only if it passes the Turing test». (Russell)
- P21:«These sharp boundaries on machine competence mean that when people talk about "machine IQ" increasing rapidly and threatening to exceed human IQ, they are talking nonsense». (Russell)
- P22:«Trying to assign an IQ to machines is like trying to get four-legged animals to compete in a human decathlon. True, horses can run fast and jump high, but they have a lot of trouble with pole-vaulting and throwing the discus». (Russell)
- P63:«for the last 20 years or so has been focused on the problems surrounding the construction of intelligent agents - systems that perceive and act in some environment» (Musk1)
- P64:«In this context, "intelligence" is related to statistical and economic notions of rationality - colloquially, the ability to make good decisions, plans, or inferences». (Musk1)
- P65:«speech recognition, image classification, autonomous vehicles, machine translation, legged locomotion, and question-answering systems». (Musk1)
- P66:«a virtuous cycle takes hold whereby even small improvements in performance are worth large sums of money, prompting greater investments in research» (Musk1)

- P67:«There is now a broad consensus that AI research is progressing steadily, and that its impact on society is likely to increase. The potential benefits are huge[...]». (Musk1)
- P68:«[...] everything that civilization has to offer is a product of human intelligence; we cannot predict what we might achieve when this intelligence is magnified by the tools AI may provide, but the eradication of disease and poverty are not unfathomable». (Musk1)
- P83:«Such decisions must not be delegated to unelected tech leaders». (Musk4)
- P84:«we call on all AI labs to immediately pause for at least 6 months the training of AI systems more powerful than GPT-4». (Musk4)
- P92:«It will change the way people work, learn, travel, get health care, and communicate with each other. Entire industries will reorient around it. Businesses will distinguish themselves by how well they use it». (Gates)
- P94:«the workforce, the legal system, privacy, bias, and more». (Gates)
- P108:«Once developers can generalize a learning algorithm and run it at the speed of a computer—an accomplishment that could be a decade away or a century away—we'll have an incredibly powerful AGI». (Gates)
- P115:«Crucially, AI is not magic, but rather just a set of engineering techniques and algorithms, each with its own strengths and weaknesses, suitable for some problems but not others». (Marcus)
- P117:«As we will discuss later, deep learning systems in no way capture the complexity and diversity of actual brains and the components of deep learning systems lack virtually all the complexity of actual neurons. As the late Francis Crick noted, it's a serious stretch to call them brain-like». (Marcus)
- P126:«Senator, I'm an AI researcher. Your description of ChatGPT is dangerously misinformed. Every sentence is incorrect. I hope you will learn more about how this system actually works, how it was trained, and what it's limitations are». (Mitchell)
- P127: «In other words, the intelligent part of your mind can't harness the fast-adding skills of your own neurons, and for good reason. This barrier — between the "self" that you are aware of and the detailed activity of your brain — permits the kind of thinking that matters for survival without getting overwhelmed ("addlebrained") by your own thought processes». (Mitchell)
- P128:«The "unexplainable" narrative gives rise to fear, and it has been argued that, to a degree, public fear of AI is actually useful for the tech companies selling it, since the

flip-side of the fear is the belief that these systems are truly powerful and big companies would be foolish not to adopt them». (Mitchell)

- P129:«Indeed, the way we talk about machine abilities influences our conceptions of how general those abilities really are». (Mitchell)
- P130:«This model views the mind as a kind of computer, which inputs, stores, processes, and outputs information. The body does not play much of a role except in the input (perception) and output (behavior) stages. Under this view, cognition takes place wholly in the brain, and is, in theory, separable from the rest of the body». (Mitchell)
- P131:«It's clear that to make and assess progress in AI more effectively, we will need to develop a better vocabulary for talking about what machines can do. And more generally, we will need a better scientific understanding of intelligence as it manifests in different systems in nature». (Mitchell)
- P159:«But just as the symbol system people are allowed to work incrementally in their goals, so should the physical grounding people be allowed». (Brooks)
- P163:«But we are not good at modelling living systems, at small or large scales. Something is wrong». (Brooks)
- P165:«We would then need to find new ways of thinking about living systems to make any progress, and this will be much more disruptive to all biology». (Brooks)
- P166:«An analogy to the sort of thing that might be missing is computation — not as the undiscovered feature itself but as an analogy for the type of thing we might be looking for». (Brooks)
- P167:«Neuroscience uses computation as a metaphor, and I question whether that's the right set of metaphors». (Brooks)
- P168:«My point is that I don't think that classical computation is the right mechanism to think about quantum mechanics. There are other metaphors». (Brooks)
- P169:«Is information processing the right metaphor there? Or are control theory and resonance and synchronization the right metaphor? We need different metaphors at different times, rather than just computation». (Brooks)
- P170:«We have fairly simple dynamics in our computational spaces because that's what we can generate with computation». (Brooks)
- P171:«The reason for why we got stuck in this cul-de-sac for so long was because Moore's law just kept feeding us, and we kept thinking, "Oh, we're making progress, we're making progress, we're making progress." But maybe we haven't been». (Brooks).

- P172:«Let me give you an example that fits your model there. We went from the Turing machine to the RAM model, and current computational complexity is really built on the RAM model of computation. It's how space and time trade off in computation». (Brooks)
- P173:«There may be something somewhat different from that that we just haven't seen yet in the large system of lots of processes happening without clear interfaces, and lots of statistical stuff going on—statistical just because you don't know everything». (Brooks)
- P174:«In short, there will be valuable tools produced, and at the same time lots of damaging misuse». (Brooks)
- P175:«GPT-n cannot reason, and it has no model of the world. It just looks at correlations between how words appear in vast quantities of text from the web, without know how they connect to the world. It doesn't even know there is a world». (Brooks)
- P176:«Many successful applications of AI have a person somewhere in the loop. Sometimes it is a person behind the scenes that the people using the system do not see, but often it is the user of the system, who provides the glue between the AI system and the real world». (Brooks)
- P177:«We tend to overestimate the effect of a technology in the short run and underestimate the effect in the long run». (Brooks)
- P178: «If I can just expand on that a little. When we see a person with some level performance at some intellectual thing, like describing what's in a picture, for instance, from that performance, we can generalize about their competence in the area they're talking about. And we're really good at that. Evolutionarily, it's something that we ought to be able to do. We see a person do something, and we know what else they can do, and we can make a judgement quickly. But our models for generalizing from a performance to a competence don't apply to AI systems». (Brooks)
- P180: «And as a species, humanity, we have changed up our mobility infrastructure multiple times. In the early 1800s, it was steam trains. We had to do enormous changes to our infrastructure. We had to put flat rails right across countries. When we started adopting automobiles around the turn from the 19th to the 20th century, we changed the roads. We changed the laws. People could no longer walk in the middle of the road like they used to. We changed the infrastructure. When you go from trains that are driven by a person to selfdriving trains, such as we see in airports and a few out there, there's a

whole change in infrastructure so that you can't possibly have a person walking on the tracks. We've tried to make this transition [to self-driving cars] without changing infrastructure. You always need to change infrastructure if you're going to do a major change». (Brooks)

- P182: «But then the magic of our robot is that it looks like a shopping cart. It's got handlebars on it. If a person goes up and grabs it, it's now a powered shopping cart or powered cart that they can move around. So [the warehouse workers] are not subject to the whims of the automation. They get to take over. When the robot's clearly doing something dumb, they can just grab it and move it, and it repairs». (Brooks)

10.2 Annex 2: Resultats SELFIE: MINMAX

CheckIn_HE_v.2021_ES - Results

Gracias por su contribución.

A continuación, encontrará la información sobre **cómo interpretar sus resultados** y justo a continuación, en la parte inferior, **su puntuación general**.

Sobre el resultado general:

Si su puntuación está por debajo de 23, es "Principiante" (A1)
Esto significa: Tiene la oportunidad de comenzar a mejorar sus habilidades con la tecnología digital. Los comentarios que obtiene de esta encuesta han identificado una serie de acciones que puede probar. Seleccione una o dos para comenzar durante el próximo periodo de enseñanza, centrándose en mejorar significativamente sus estrategias docentes. Mientras lo hace, se encontrará avanzando en el siguiente paso de la competencia digital, el nivel "Explorador".

Si su puntuación está entre 23 y 38, es "Explorador" (A2)
Esto significa: Es consciente del potencial de las tecnologías digitales y está interesado en explorarlas para mejorar la práctica pedagógica y profesional. Ha comenzado a utilizar tecnologías digitales en algunas áreas y se beneficiará de una práctica más consistente. Puede aumentar su competencia colaborando e intercambiando con compañeros, y ampliando aún más su repertorio de prácticas y habilidades digitales. Esto le llevará al siguiente paso de la competencia digital, el nivel "Integrador".

Si su puntuación está entre 39 y 56, es "Integrador" (B1)
Esto significa: Experimenta con tecnologías digitales en diferentes contextos y para diversos propósitos, integrándolos en muchas de sus prácticas docentes. Los usa creativamente para mejorar diferentes aspectos de su compromiso profesional. Está impaciente por ampliar su repertorio de prácticas. Se beneficiará al aumentar la comprensión sobre qué herramientas funcionan mejor en qué situaciones y sobre cómo adaptar las tecnologías digitales a las estrategias y métodos pedagógicos. Trate de darse más tiempo para la reflexión y la adaptación, complementado con el intercambio de estímulos colaborativos y de conocimientos, para llegar al siguiente paso, "Experto" (B2).

Si su puntuación está entre 57 y 74, es "Experto" (B2)
Esto significa: Utiliza diversas tecnologías digitales con confianza, creatividad y crítica para mejorar sus actividades profesionales. Selecciona con un propósito concreto tecnologías digitales para situaciones determinadas, y trata de entender los beneficios y desventajas de diferentes estrategias digitales. Es curioso y está abierto a nuevas ideas, sabiendo que hay muchas opciones que aún no ha probado. Utiliza la experimentación como un medio para expandir, estructurar y consolidar su repertorio de estrategias. Comparta su experiencia con otros [educadores] y

continúe desarrollando críticamente sus estrategias digitales para alcanzar el nivel “Líder” (C1).

Si su puntuación está entre 75 y 91, es “Líder” (C1)
Esto significa: Tiene un enfoque consistente y completo para aplicar las tecnologías digitales para mejorar las prácticas pedagógicas y profesionales. Confía en que cuenta con un amplio repertorio de estrategias digitales, de las que sabe cómo elegir la más adecuada para cada situación. Continuamente reflexiona y desarrolla sus prácticas. Al intercambiar experiencias con sus compañeros, se mantiene actualizado en nuevos desarrollos e ideas y ayuda a otros [educadores] a aprovechar el potencial de las tecnologías digitales para mejorar la enseñanza y el aprendizaje. Si está listo para experimentar un poco más, podrá alcanzar la última etapa de competencia, como “Pionero”.

Si su puntuación es superior a 92, es “Pionero” (C2)
Esto significa: Cuestiona la adecuación de las prácticas contemporáneas digitales y pedagógicas, en las que es líder. Le preocupan las limitaciones o inconvenientes de estas prácticas y se siente motivado por el impulso de innovar aún más la educación. Experimenta con tecnologías digitales altamente innovadoras y complejas y/o desarrolla enfoques pedagógicos novedosos. Dirige la innovación y es un modelo a seguir para otros profesores. Para comprender mejor su perfil competencial, debe observar su desempeño por área. Debido al número limitado de elementos utilizados en esta herramienta, desafortunadamente no es posible calcular una puntuación fiable por área.

Sin embargo, para ofrecerle una **primera aproximación para ayudarle a conocer sus debilidades y fortalezas relativas por área**, se aplican las siguientes reglas generales:

En las Áreas 1 y 3:
Principiante/Explorador (A): inferior a 8 puntos
Integrador/Experto (B): 8-13 puntos
Líder/Pionero (C): más de 13 puntos

En las Áreas 2, 4, 5 y 7:
Principiante/Explorador (A): inferior a 6 puntos
Integrador/Experto (B): 6-9 puntos
Líder/Pionero (C): más de 9 puntos

En el Área 6:
Principiante/Explorador (A): inferior a 9 puntos
Integrador/Experto (B): 9-16 puntos
Líder/Pionero (C): más de 16 puntos

Summary:

Your Score 0

Maximum Score 150

Section	Score for this Section	
rea 1: Compromiso profesional	0/24	
rea 2: Contenidos digitales	0/18	
rea 3: Enseanza y aprendizaje	0/24	
rea 4: Evaluacin y retroalimentacin	0/18	
rea 5: Empoderamiento de los estudiantes	0/18	
rea 6: Desarrollo de la competencia digital de los estudiantes	0/30	
rea 7: Educacin abierta (basada en el marco OpenEdu)	0/18	

Scores by Question:

Código de participación

Your answer ✔ no answer given

0 out of 0 points

País

Your answer ✖ Andorra

0 out of 0 points

Dedicación

Your answer

0 out of 0 points

Categoria profesional

Your answer

0 out of 0 points

¿Cómo evalúa actualmente su competencia digital como profesor? Asigne un nivel de competencia de A1 a C2, en el que A1 es el más bajo y C2 el más alto. Probablemente soy:

Your answer ✔ C1: Líder

0 out of 0 points

rea 1: Compromiso profesional

Score for this Section: 0/24

Utilizo diferentes canales digitales para mejorar la comunicación con los estudiantes y compañeros cuando es necesario. Por ejemplo: correos electrónicos, blogs, el sitio web de la organización educativa, sistema de gestión del aprendizaje –LMS–, apps, etc.

Your answer ✔ No uso canales de comunicación digital

El uso de canales de comunicación digital puede ayudarle a optimizar sus contactos con los estudiantes y compañeros. Empiece escribiendo correos electrónicos o configure un blog de curso para intercambiar información.

[Para subir de nivel]: Intente comunicarse por correo electrónico o un sistema de mensajería instantánea.

0 out of 6 points

Uso tecnologías digitales cuando es necesario para trabajar junto a otros compañeros dentro y fuera de mi organización educativa

Your answer

✔ No colaboro con otros profesores

Si aún no existe una cultura de colaboración en su organización, una opción para comenzar podría ser, por ejemplo, ofrecerse a compartir sus materiales e ideas con sus compañeros y pedirles que compartan sus materiales con usted. Además, unirse a una comunidad profesional en línea le permite inspirarse en los materiales que han creado otros profesores en su país, en toda Europa y en cualquier lugar del mundo. Si comparte sus planes de estudio y materiales con ellos, puede obtener sus comentarios e ideas sobre cómo adaptarlos a diferentes situaciones o cómo mejorarlos. Tal intercambio es a menudo una experiencia enriquecedora, a nivel personal y profesional.

[Para subir de nivel]: Intercambie información en colaboración con sus compañeros.

0 out of 6 points

Desarrollo activamente mi competencia digital para la docencia

Your answer

✔ No trabajo en el desarrollo de mi competencia digital para la docencia

Muchos profesores encuentran que les falta tiempo y apoyo suficiente para el desarrollo profesional. Sin embargo, hay formas en las que puede trabajar para mejorar sus competencias para la docencia digital sin invertir mucho tiempo extra. Un primer paso podría ser participar en una práctica reflexiva y preguntarse después de cada clase: ¿He utilizado las tecnologías digitales con un valor añadido? ¿Qué he logrado con ellas que no podría haber logrado de otra manera? ¿Qué puedo cambiar para mejorar la correspondencia entre la tecnología que seleccioné y los objetivos de aprendizaje establecidos? Intente identificar qué factores han contribuido a las coincidencias positivas y negativas entre las herramientas digitales y los resultados de aprendizaje, y piense cómo mejorar esta comparación.

[Para subir de nivel]: Reflexione sobre su docencia digital como una rutina diaria.

0 out of 6 points

Participo en cursos de formación en línea cuando se presenta la oportunidad Por ejemplo: cursos en línea, MOOC, seminarios web o conferencias virtuales

Your answer

✔ Esto es algo que todavía no he considerado

Los numerosos recursos disponibles en Internet pueden facilitar la actualización de sus habilidades independientemente de la ubicación y el tiempo, especialmente si no cuenta con tiempo suficiente para participar en actividades de desarrollo profesional continuo más formal. Una opción, para empezar, podría ser pensar en una palabra de moda en la teoría pedagógica contemporánea (como "aula invertida") o algún enfoque que le guste a un compañero y del que sabe muy poco. Una búsqueda en Internet le proporcionará una serie de vídeos, discusiones o blogs, que le facilitarán más hilos y enlaces para el seguimiento. Siguiendo estos hilos y enlaces, aprenderá mucho sobre este concepto y se dará cuenta de dónde profundizar, en caso de que lo desee. Sin darse cuenta, ha "participado en cursos de formación en línea cuando se presenta la oportunidad".

[Para subir de nivel]: Busque en Internet una estrategia de enseñanza sobre la que le gustaría aprender más.

0 out of 6 points

rea 2: Contenidos digitales

Score for this Section: 0/18

Utilizo diferentes sitios de Internet y estrategias de búsqueda para encontrar y seleccionar diferentes recursos digitales

Your answer

✔ No sé cómo usar Internet para buscar recursos útiles

Aunque cree que Internet puede ser útil para encontrar nuevos recursos para la docencia, todavía no ha desarrollado las habilidades necesarias para utilizar los mecanismos de búsqueda en línea.
[Para subir de nivel]: Solicite ayuda e intente buscar materiales didácticos adecuados en línea.

0 out of 6 points

Creo mis propios contenidos digitales y modifico otros existentes para adaptarlos a mis necesidades

Your answer

✔ No creo mis propios recursos digitales

Siente que le faltan algunas habilidades para crear sus propios recursos digitales, por lo que prefiere reutilizar los de sus compañeros.

[Para subir de nivel]: Experimente creando sus propios recursos digitales.

0 out of 6 points

Protejo de forma efectiva los datos personales como, por ejemplo, exámenes, calificaciones o datos personales

Your answer

✔ No necesito hacerlo porque la institución en la que trabajo se encarga de ello

La mayoría de las instituciones tienen políticas de protección de datos. Sin embargo, necesita seguirlas para que cumplan su función. Asegúrese de usar contraseñas que no se puedan adivinar fácilmente y evite que otras personas vean cómo las utiliza. Cambie sus contraseñas regularmente y elimine los datos que ya no necesita como, por ejemplo, los datos personales de sus antiguos estudiantes. Proteja sus dispositivos personales si tiene datos en ellos. Use el cifrado cuando comparta con otros colegas archivos que tengan datos personales.

[Para subir de nivel]: Revise con atención cómo comparte archivos y cómo protege sus dispositivos personales.

0 out of 6 points

rea 3: Enseanza y aprendizaje

Score for this Section: 0/24

Valoro con atención cómo, cuándo y por qué usar tecnologías digitales en el aula con mis estudiantes, para garantizar que aporten valor añadido

Your answer	✔ No uso o uso esporádicamente tecnología en mis clases Hay varias formas de iniciarse. Es muy probable que todos sus estudiantes tengan un dispositivo digital con ellos, aunque solo sea un teléfono móvil. Si su institución permite el uso en el aula de dispositivos móviles, puede diversificar su docencia con tareas prácticas para que realicen, por ejemplo, cuestiones para buscar o calcular, rellenar pequeñas encuestas y cuestionarios, etc. La ventaja de esto es que puede involucrar de forma activa a los estudiantes en clase, aumentando su aprendizaje. Además, le permite recopilar evidencias sobre qué materias de su docencia son entendidas suficientemente por los estudiantes y cuáles deberá repasar. [Para subir de nivel]: Pida a sus estudiantes que usen dispositivos digitales para pequeñas actividades en el aula.	0 out of 6 points	

Superviso las actividades e interacciones de mis estudiantes en los entornos colaborativos en línea que utilizamos

Your answer	✔ No uso entornos digitales con mis estudiantes Para obtener más información acerca de sus estudiantes y sus necesidades de aprendizaje, considere involucrarlos en tareas de trabajo en grupo. El trabajo grupal puede promover el aprendizaje y, si se utilizan entornos digitales, es mucho más fácil dar el apoyo que su estudiante necesita. Los entornos colaborativos en línea pueden ayudar a canalizar la comunicación con sus estudiantes (por ejemplo, resolviendo los problemas y preguntas que tienen) y la colaboración (p. ej., proyectos en los que trabajan en grupos). Existen muchos servicios independientes o integrados que abordan estos dos objetivos de manera conjunta o por separado. Muchos de ellos son de código abierto o están disponibles de forma gratuita. Para comenzar, pregunte a sus colegas si pueden recomendarle una solución concreta o busque recomendaciones en Internet. Pruebe qué solución funciona mejor para usted. [Para subir de nivel]: Pruebe un entorno colaborativo en línea.	0 out of 6 points	

Cuando mis estudiantes trabajan en grupo, utilizan tecnologías digitales para adquirir y plasmar los conocimientos

Your answer

✔ No sé cómo integrar las tecnologías digitales en actividades de aprendizaje colaborativo

Hoy en día, el trabajo y la investigación se basan esencialmente en procesos colaborativos. Para preparar a los estudiantes para esta realidad, es importante integrar los procesos de colaboración en sus estudios. Por esta razón, el trabajo en grupo debe ser una parte integral de la Educación Superior. Un entorno de aprendizaje digital colaborativo, como una wiki o un blog, puede ayudar al trabajo de los estudiantes en grupo a estructurar su colaboración y plasmar de forma efectiva su aprendizaje. Como profesor, debe intentar identificar las situaciones de aprendizaje colaborativo que se beneficiarían del uso de la tecnología.

[Para subir de nivel]: Implemente actividades de aprendizaje colaborativo con el apoyo de tecnologías digitales siempre que sea apropiado.

0 out of 6 points

Utilizo tecnologías digitales para permitir a mis estudiantes planificar, documentar y monitorizar su propio proceso de aprendizaje Por ejemplo: autoevaluaciones, portafolios digitales para documentar y exponer, diarios/blogs en línea para reflexiones, etc.

Your answer

✔ No es posible en mi entorno de trabajo

Para empezar a utilizar herramientas digitales para evaluar, considere la posibilidad de integrar un breve cuestionario o encuesta en sus cursos en línea o, en un aula física, como una actividad al final de cada tema o unidad. Otra opción podría ser introducir un diario de aprendizaje en línea, p. ej., en forma de blog, donde los estudiantes documenten y reflexionen sobre sus logros y necesidades de aprendizaje.

[Para subir de nivel]: Comience con una encuesta rápida al final de cada tema.

0 out of 6 points

rea 4: Evaluacin y retroalimentacin

Score for this Section: 0/18

Uso herramientas digitales de evaluación para monitorizar el progreso de los estudiantes

Your answer

✔ No sigo el progreso de los estudiantes con medios digitales

Para comprender lo que sus estudiantes han aprendido y lo que aún no han entendido bien, debería monitorizar continuamente su progreso -sea esta o no una práctica común en su institución-. La forma más fácil de hacerlo es crear una pequeña prueba o juego en cada unidad, o una actividad para casa, para que usted y sus estudiantes puedan tener una visión de lo que necesita ser revisado y lo que el estudiante ha entendido. Posteriormente puede adaptar su docencia a estos resultados.
[Para subir de nivel]: Explore los cuestionarios digitales.

0 out of 6 points

Analizo todos los datos disponibles para identificar de manera efectiva a los estudiantes que necesitan apoyo adicional Nota: "Datos" incluye: información personal, actividades de participación de los estudiantes, información sobre el rendimiento, calificaciones, asistencia e interacciones sociales en entornos (en línea); "Los estudiantes que necesitan apoyo adicional" son: estudiantes que están en riesgo de abandonar o tener un bajo rendimiento; estudiantes que tienen trastornos de aprendizaje o *necesidades educativas especiales; o estudiantes que carecen de habilidades transversales (p. ej., habilidades sociales, verbales o de estudio).

Your answer

✔ La información de este tipo de estudiantes no está disponible para mí y /u otra persona de mi institución la analiza

Es importante crear un ambiente de aprendizaje en el que los estudiantes que tienen necesidades educativas especiales o que necesiten apoyo adicional se sientan cómodos compartiendo esta información con usted. Los estudiantes que han estado desconectados de la educación reglada a menudo se sienten abrumados por el ritmo y el formato de estudio, especialmente en los cursos en línea. Estar atento a los signos de desvinculación le ayudará a identificar a los estudiantes en riesgo y a ayudarlos a volver a la normalidad.
[Para subir de nivel]: Examine los datos disponibles para identificar a los estudiantes que tienen dificultades.

0 out of 6 points

Uso tecnologías digitales para proporcionar retroalimentación a los estudiantes

Your answer

✔ La retroalimentación no es necesaria en mi entorno de trabajo

Uno de los propósitos principales de la evaluación es indicar las áreas en las que necesitan mejorar los estudiantes. La retroalimentación es esencial para que los estudiantes puedan entender cómo pueden mejorar.

[Para subir de nivel]: Proporcione a los estudiantes retroalimentación sobre su proceso de aprendizaje y resultados.

0 out of 6 points

rea 5: Empoderamiento de los estudiantes

Score for this Section: 0/18

Cuando creo tareas digitales para los estudiantes, considero y abordo posibles dificultades prácticas o técnicas Por ejemplo: acceso igualitario a dispositivos y recursos digitales; problemas de interoperabilidad y conversión; falta de habilidades digitales

Your answer

✔ No creo tareas digitales

Para probar las tareas digitales, considere pedir a los estudiantes que busquen en línea información relevante para el tema de estudio y que presenten sus hallazgos en formato digital. Pregúnteles sobre los problemas que tuvieron con esta tarea y ajuste las reglas (p. ej. fechas límite, formato de presentación) para permitir que todos los estudiantes participen en las tareas digitales.
[Para subir de nivel]: Explore tareas digitales.

0 out of 6 points

Utilizo tecnologías digitales para ofrecer a los estudiantes opciones de aprendizaje personalizadas Por ejemplo: planteo diferentes tareas digitales a los estudiantes para abordar las necesidades de aprendizaje individuales, preferencias e intereses

Your answer	✔ En mi entorno laboral, todos los estudiantes están obligados a hacer las mismas actividades, independientemente de su nivel Aunque todos los estudiantes están obligados a realizar las mismas actividades, debe considerar qué estudiantes necesitan apoyo adicional y a cuáles debe estimular más. Tratar a los estudiantes por igual no significa ofrecerles a todos el mismo tratamiento, sino ofrecerles a cada uno el tratamiento que necesiten, especialmente si todos tienen que alcanzar el mismo objetivo de aprendizaje al final. Combinar diferentes estrategias de enseñanza y aprendizaje e implementar una diversidad de actividades de aprendizaje puede dar como resultado un aprendizaje más efectivo y más profundo para todos los estudiantes. [Para subir de nivel]: Aborde diferentes *necesidades de aprendizaje y preferencias al enseñar.	0 out of 6 points	

Uso tecnologías digitales para que los estudiantes participen activamente en clase o en línea

Your answer

✔ En mi lugar de trabajo no es posible involucrar activamente a los estudiantes en clase o en línea

Incluso si su aula no está equipada digitalmente en el campus, la mayoría de sus estudiantes probablemente tienen un dispositivo digital con acceso a Internet. Comience con pedir a los estudiantes que busquen información en Internet como tarea en casa. O pídales que tomen fotos o vídeos que ejemplifiquen el tema de estudio. En clase, los estudiantes pueden reunir la información que encontraron, debatirla en grupos pequeños y convertirla en una presentación o artefacto.

Si piensa que este tipo de trabajo no es lo que se espera de sus estudiantes en el currículum, vuelva a leer cuidadosamente los requisitos curriculares de su plan de estudios y hable con sus asesores. Encontrará que hay más espacio para la creatividad de lo que pensaba.

[Para subir de nivel]: Comience y haga que sus estudiantes se involucren.

0 out of 6 points

rea 6: Desarrollo de la competencia digital de los estudiantes

Score for this Section: 0/30

Enseño a los estudiantes cómo evaluar la fiabilidad de la información

Your answer

✔ Esto no es posible en mi asignatura o lugar de trabajo

Es cierto que la adquisición de competencias para el tratamiento de la información es más relevante para algunas materias que para otras. Sin embargo, incluso si su asignatura es, p. ej. matemáticas, debe permitir que sus estudiantes busquen información y materiales de aprendizaje en línea y puedan juzgar lo bueno de lo malo y la información correcta de la incorrecta.

Para abordar de manera significativa la adquisición de competencias para el tratamiento de la información en su materia, puede, por ejemplo, integrarla en una actividad de revisión: presente a los estudiantes un sitio web o contenido audiovisual tomado de Internet sobre un tema que acaban de estudiar y pídales que identifiquen inexactitudes, información que falte o sea sesgada.

[Para subir de nivel]: Utilice una fuente de información con fallos en una actividad de revisión para fomentar la evaluación de la información.

0 out of 6 points

Configuro tareas que requieren que los estudiantes usen medios digitales para comunicarse y colaborar entre sí o con una audiencia externa

Your answer	✔ Esto no es posible en mi asignatura o lugar de trabajo La comunicación digital es una habilidad básica importante en nuestras sociedades. Es responsabilidad de todas las instituciones educativas, en todos los niveles, desarrollar esta habilidad en los estudiantes. Para alentar a los alumnos a comunicarse entre sí, puede ayudar a crear una comunidad o grupo en un entorno de colaboración en línea y establecer una tarea colaborativa concreta que resolver utilizando este entorno. Para animar a los estudiantes a comunicarse con una audiencia externa, una actividad basada en una entrevista puede servir como punto de partida. Cualquiera que sea la tarea concreta en cuestión, anime a los estudiantes a descubrir y desarrollar colaborativamente reglas efectivas para la comunicación y la colaboración. Motívelos a documentar sus reglas y fortalecerlas entre ellos. Desafíe sus reglas mediante la integración de tareas o variaciones que requieran diferentes estrategias o normas de colaboración para la comunicación. [Para subir de nivel]: Establezca incentivos para la comunicación y colaboración.	0 out of 6 points	

Configuro tareas que requieran a los estudiantes crear contenido digital Por ejemplo: vídeos, audios, fotos, presentaciones digitales, blogs o wikis

Your answer

✔ No sé cómo hacerlo

Es cierto que en algunas asignaturas es más fácil que en otras integrar actividades digitales para los estudiantes. Sin embargo, cuando lo analice, encontrará alguna unidad en la que los estudiantes podrían crear contenido, por ejemplo, realizando una entrevista y grabándola, haciendo fotos de ejemplos para estudiar, escribiendo un texto y publicándolo en formato en línea, diseñando un artefacto digital con un *software* que utilicen...

De esta manera, motiva a sus estudiantes para que trabajen y estudien su materia, aumenta su participación activa en el proceso de aprendizaje y también promueve sus habilidades para crear contenido digital.

[Para subir de nivel]: Integre actividades digitales.

0 out of 6 points

Enseño a los estudiantes a usar la tecnología digital de manera segura y responsable

Your answer

✔ Esto no es posible en mi asignatura o lugar de trabajo

Incluso si no prevé actividades de aprendizaje que requieran que los estudiantes usen Internet, estos a menudo usan información en línea y estrategias de comunicación para complementar su aprendizaje. Necesitan comprender su huella digital, cómo proteger su identidad digital y cómo evitar la divulgación de información personal.

Para garantizar que los estudiantes estén al tanto de las reglas de protección de datos existentes, puede ser útil resumir la normativa en forma de una guía del curso.

[Para subir de nivel]: Debata las reglas de comunicación en línea con sus estudiantes.

0 out of 6 points

Animo a los estudiantes a utilizar las tecnologías digitales de manera creativa para resolver problemas concretos Por ejemplo, superar obstáculos o retos emergentes en el proceso de aprendizaje

Your answer	✔ Esto no es posible con mis estudiantes, en mi lugar de trabajo Es importante que los estudiantes puedan formular sus problemas para planificar su aprendizaje, comunicar sus ideas o comprender el contenido del curso; identificar las barreras concretas encontradas; y animarles a que piensen en las formas de superarlos. Para usted como profesor esto significa que debe estar abierto a las diferentes maneras en que el estudiante supera los obstáculos. Y esto implica que debe tratar de fomentar esta forma de encontrar soluciones aunque para usted puedan parecer ineficientes, arbitrarias, científicamente dudosas o en otros aspectos poco ortodoxas. Usted puede, y debe, animar a los estudiantes a trabajar en los defectos de sus estrategias de apropiación, mientras se dan cuenta y valoran que dieron el primer paso para superar un obstáculo importante para su aprendizaje. [Para subir de nivel]: Anime a los estudiantes a superar de manera creativa los desafíos de la comunicación.	0 out of 6 points	

rea 7: Educacin abierta (basada en el marco OpenEdu)

Score for this Section: 0/18

Sé cómo encontrar y utilizar licencias abiertas en recursos educativos

Your answer

✔ No sé qué es un Recurso Educativo Abierto (REA)

Es posible que haya oído hablar de los Recursos Educativos Abiertos (REA) pero no está seguro de lo que significa. Puede pensar que todos los recursos disponibles en línea pueden ser utilizados y compartidos siempre que sean gratuitos. Tenga en cuenta que un recurso educativo sin una licencia abierta no es un recurso educativo abierto, incluso si el recurso está disponible en línea y es gratuito.

[Para subir de nivel]: Póngase al día con la definición de un Recurso Educativo Abierto consultando la "dimensión de contenido" del Marco OpenEdu y las Guías prácticas sobre Educación en abierto para académicos (JRC 2016, 2019).

0 out of 6 points

Adopto prácticas educativas abiertas en mi docencia para hacerla más inclusiva

Your answer

✔ No sé aplicar prácticas educativas abiertas en mi docencia

Ya sea porque no sabe qué son las Prácticas Educativas Abiertas (OEP), o carece de habilidades digitales para crear, reutilizar y publicar materiales educativos como REA o, incluso, porque tiene inseguridades con respecto a llegar a una audiencia diversa más allá de los muros institucionales, debe saber que no está solo y que éstas son prácticas desafiantes para la mayoría de los profesores. Sin embargo, una vez que empiece a romper estas barreras, verá que las recompensas merecen la pena. Intente alcanzar una mejor comprensión de las prácticas educativas abiertas. Considere cómo su institución podría respaldarlo y pedir el apoyo de compañeros que ya están en este camino.

[Para subir de nivel]: Actualízate con la definición de prácticas educativas abiertas comprobando la dimensión "pedagogía" del Marco OpenEdu (JRC, 2016) y las Directrices prácticas sobre educación abierta para académicos.

0 out of 6 points

Publico mi investigación en revistas científicas abiertas, así como mis datos de investigación siempre que sea posible

Your answer

✔ No estoy familiarizado con el concepto “Ciencia Abierta”

Es posible que haya oído hablar de prácticas científicas abiertas, pero no se ha familiarizado con lo que son. No sabe cómo podría aplicarse a la investigación que produce.

[Para subir de nivel]: Póngase al día con las formas en que puede abrir el acceso a su propia docencia e investigación consultando la "dimensión de investigación" de las Directrices prácticas sobre educación abierta para académicos (JRC 2019).

0 out of 6 points

Es usted…

Your answer

✔ Hombre

0 out of 0 points

¿Cuál es su edad?

Your answer

✔ Menos de 25

0 out of 0 points

¿En cuál de las siguientes ramas desempeña su docencia?

Your answer

✔ Artes y Humanidades

0 out of 0 points

Incluyendo este curso académico, ¿Cuántos años lleva trabajando como profesor?

Your answer

✔ 1-5

0 out of 0 points

¿En qué porcentaje de sus clases ha utilizado tecnología o herramientas digitales en los últimos tres meses?

Your answer ✔ 0-10%

0 out of 0 points

¿Cuánto tiempo lleva utilizando tecnología en sus clases?

Your answer ✔ Todavía no he utilizado tecnología en mis clases

0 out of 0 points

¿Qué herramientas digitales ha utilizado usted o sus estudiantes para enseñar y aprender? Puede seleccionar varias opciones

Your answer ✔ Presentaciones
✖ Not all correct answers selected

0 out of 0 points

¿Qué edades tienen los estudiantes a los que enseña?

Your answer ✔ Menos de 18
✖ Not all correct answers selected

0 out of 0 points

¿Cuál es el perfil principal de sus estudiantes? Puede seleccionar varias opciones

Your answer ✔ Pregrado
✖ Not all correct answers selected

0 out of 0 points

¿Cómo evalúa su competencia digital docente como profesor ahora, después de responder al cuestionario? Asigne un nivel de competencia de A1 a C2, donde A1 es el más bajo y C2 el más alto Probablemente soy:

Your answer ✔ A1: Principiante

0 out of 0 points

Contact	jrc-digcompedu@ec.europa.eu
Useful links	DigCompEdu OpenEdu
Background Documents	Glosario Guia Educación Abierta

Contribution ID	66f35070-ef3f-4ee6-a005-e666f0585f3c
Completed at	18/08/2024 18:31:58
Completion time	-

CheckIn_HE_v.2021_ES - Results

Gracias por su contribución.

A continuación, encontrará la información sobre **cómo interpretar sus resultados** y justo a continuación, en la parte inferior, **su puntuación general**.

Sobre el resultado general:

Si su puntuación está por debajo de 23, es "Principiante" (A1)
Esto significa: Tiene la oportunidad de comenzar a mejorar sus habilidades con la tecnología digital. Los comentarios que obtiene de esta encuesta han identificado una serie de acciones que puede probar. Seleccione una o dos para comenzar durante el próximo periodo de enseñanza, centrándose en mejorar significativamente sus estrategias docentes. Mientras lo hace, se encontrará avanzando en el siguiente paso de la competencia digital, el nivel "Explorador".

Si su puntuación está entre 23 y 38, es "Explorador" (A2)
Esto significa: Es consciente del potencial de las tecnologías digitales y está interesado en explorarlas para mejorar la práctica pedagógica y profesional. Ha comenzado a utilizar tecnologías digitales en algunas áreas y se beneficiará de una práctica más consistente. Puede aumentar su competencia colaborando e intercambiando con compañeros, y ampliando aún más su repertorio de prácticas y habilidades digitales. Esto le llevará al siguiente paso de la competencia digital, el nivel "Integrador".

Si su puntuación está entre 39 y 56, es "Integrador" (B1)
Esto significa: Experimenta con tecnologías digitales en diferentes contextos y para diversos propósitos, integrándolos en muchas de sus prácticas docentes. Los usa creativamente para mejorar diferentes aspectos de su compromiso profesional. Está impaciente por ampliar su repertorio de prácticas. Se beneficiará al aumentar la comprensión sobre qué herramientas funcionan mejor en qué situaciones y sobre cómo adaptar las tecnologías digitales a las estrategias y métodos pedagógicos. Trate de darse más tiempo para la reflexión y la adaptación, complementado con el intercambio de estímulos colaborativos y de conocimientos, para llegar al siguiente paso, "Experto" (B2).

Si su puntuación está entre 57 y 74, es "Experto" (B2)
Esto significa: Utiliza diversas tecnologías digitales con confianza, creatividad y crítica para mejorar sus actividades profesionales. Selecciona con un propósito concreto tecnologías digitales para situaciones determinadas, y trata de entender los beneficios y desventajas de diferentes estrategias digitales. Es curioso y está abierto a nuevas ideas, sabiendo que hay muchas opciones que aún no ha probado. Utiliza la experimentación como un medio para expandir, estructurar y consolidar su repertorio de estrategias. Comparta su experiencia con otros [educadores] y

continúe desarrollando críticamente sus estrategias digitales para alcanzar el nivel “Líder” (C1).

Si su puntuación está entre 75 y 91, es “Líder” (C1)
Esto significa: Tiene un enfoque consistente y completo para aplicar las tecnologías digitales para mejorar las prácticas pedagógicas y profesionales. Confía en que cuenta con un amplio repertorio de estrategias digitales, de las que sabe cómo elegir la más adecuada para cada situación. Continuamente reflexiona y desarrolla sus prácticas. Al intercambiar experiencias con sus compañeros, se mantiene actualizado en nuevos desarrollos e ideas y ayuda a otros [educadores] a aprovechar el potencial de las tecnologías digitales para mejorar la enseñanza y el aprendizaje. Si está listo para experimentar un poco más, podrá alcanzar la última etapa de competencia, como “Pionero”.

Si su puntuación es superior a 92, es “Pionero” (C2)
Esto significa: Cuestiona la adecuación de las prácticas contemporáneas digitales y pedagógicas, en las que es líder. Le preocupan las limitaciones o inconvenientes de estas prácticas y se siente motivado por el impulso de innovar aún más la educación. Experimenta con tecnologías digitales altamente innovadoras y complejas y/o desarrolla enfoques pedagógicos novedosos. Dirige la innovación y es un modelo a seguir para otros profesores. Para comprender mejor su perfil competencial, debe observar su desempeño por área. Debido al número limitado de elementos utilizados en esta herramienta, desafortunadamente no es posible calcular una puntuación fiable por área.

Sin embargo, para ofrecerle una **primera aproximación para ayudarle a conocer sus debilidades y fortalezas relativas por área**, se aplican las siguientes reglas generales:

En las Áreas 1 y 3:
Principiante/Explorador (A): inferior a 8 puntos
Integrador/Experto (B): 8-13 puntos
Líder/Pionero (C): más de 13 puntos

En las Áreas 2, 4, 5 y 7:
Principiante/Explorador (A): inferior a 6 puntos
Integrador/Experto (B): 6-9 puntos
Líder/Pionero (C): más de 9 puntos

En el Área 6:
Principiante/Explorador (A): inferior a 9 puntos
Integrador/Experto (B): 9-16 puntos
Líder/Pionero (C): más de 16 puntos

Summary:

Your Score 150

Maximum Score 150

Section	Score for this Section	
rea 1: Compromiso profesional	24/24	
rea 2: Contenidos digitales	18/18	
rea 3: Enseanza y aprendizaje	24/24	
rea 4: Evaluacin y retroalimentacin	18/18	
rea 5: Empoderamiento de los estudiantes	18/18	
rea 6: Desarrollo de la competencia digital de los estudiantes	30/30	
rea 7: Educacin abierta (basada en el marco OpenEdu)	18/18	

Scores by Question:

Código de participación

Your answer ✔ no answer given

0 out of 0 points

País

Your answer ✖ Andorra

0 out of 0 points

Dedicación

Your answer ✔ Prefiero no contestar

0 out of 0 points

Categoria profesional

Your answer ✔ Prefiero no contestar

0 out of 0 points

¿Cómo evalúa actualmente su competencia digital como profesor? Asigne un nivel de competencia de A1 a C2, en el que A1 es el más bajo y C2 el más alto. Probablemente soy:

Your answer ✔ C2: Pionero

0 out of 0 points

rea 1: Compromiso profesional

Score for this Section: 24/24

Utilizo diferentes canales digitales para mejorar la comunicación con los estudiantes y compañeros cuando es necesario. Por ejemplo: correos electrónicos, blogs, el sitio web de la organización educativa, sistema de gestión del aprendizaje –LMS–, apps, etc.

Your answer ✔ Planeo con confianza y adapto mi estrategia de comunicación digital utilizando una variedad de tecnologías digitales para satisfacer mis necesidades comunicativas en el contexto de mis interlocutores

Usted se siente seguro utilizando tecnologías digitales para comunicarse con estudiantes y compañeros. Puede planificar sus necesidades de comunicación utilizando una variedad de tecnologías, teniendo en cuenta los diferentes contextos y los resultados de comunicación esperados derivados del tipo de tecnología elegida.

6 out of 6 points

Uso tecnologías digitales cuando es necesario para trabajar junto a otros compañeros dentro y fuera de mi organización educativa

Your answer

✔ En conjunto, creo, reutilizo y comparto materiales con otros profesores y estudiantes en una red en línea

No solo crea nuevos materiales educativos conjuntamente con otros compañeros en línea, sino que también comparte sus propios materiales y reutiliza los materiales que han compartido con usted.

6 out of 6 points

Desarrollo activamente mi competencia digital para la docencia

Your answer

✔ Dirijo la innovación docente utilizando tecnologías digitales en mi institución

Tiene competencias avanzadas en el uso de tecnologías digitales para la docencia, por lo que fomenta la innovación a nivel organizativo. Si bien es importante que cada uno de ustedes continúe trabajando en sus fortalezas y debilidades individuales y aprendan unos de otros, es igual de importante debatir cómo toda la organización puede beneficiarse de sus estrategias de enseñanza innovadoras y hacer propuestas concretas para una estrategia de innovación a nivel institucional. No importa si no todas sus propuestas tienen éxito, lo importante es que la institución, en su conjunto, tome conciencia del potencial que tiene tanto usted como sus compañeros y lo aproveche para innovar en la enseñanza y el aprendizaje en toda la organización.

6 out of 6 points

Participo en cursos de formación en línea cuando se presenta la oportunidad Por ejemplo: cursos en línea, MOOC, seminarios web o conferencias virtuales

Your answer

✔ Estoy acreditado profesionalmente en el uso de diferentes tecnologías para la enseñanza y el aprendizaje

Esto significa que no solo utiliza diferentes tecnologías y enseña a otros, sino que también ha obtenido una certificación profesional de sus habilidades digitales.

6 out of 6 points

rea 2: Contenidos digitales

Score for this Section: 18/18

Utilizo diferentes sitios de Internet y estrategias de búsqueda para encontrar y seleccionar diferentes recursos digitales

Your answer

✔ No solo busco y selecciono diferentes recursos digitales, sino que también asumo el liderazgo en el fomento del uso de los mismos en mi institución

Esto significa que no solo puede crear y utilizar recursos digitales, sino que también desempeña un papel activo en su institución para ayudar a otros a beneficiarse y utilizar los recursos digitales.

6 out of 6 points

Creo mis propios contenidos digitales y modifico otros existentes para adaptarlos a mis necesidades

Your answer

✔ Adapto, uso, comparto e, incluso, creo recursos interactivos más complejos, como vídeos, pruebas de opción múltiple en línea, aplicaciones de realidad virtual, etc.

Lo importante para usted, en este nivel tan elevado, es recordar que la tecnología es un medio y no un fin. Cuando combine las diferentes características de las diversas herramientas, programas y aplicaciones digitales que utiliza, mantenga su enfoque en los objetivos de aprendizaje concretos y en las necesidades y preferencias de sus estudiantes.

6 out of 6 points

Protejo de forma efectiva los datos personales como, por ejemplo, exámenes, calificaciones o datos personales

Your answer

✔ Protejo los datos digitales y aplico el RGPD (Reglamento General de Protección de Datos) cuando se trata de temas identificables, como los datos relacionados con mis estudiantes

No solo utiliza técnicas de protección de datos digitales, sino que también busca mantenerse actualizado con las últimas regulaciones sobre protección de datos, entendiendo así el Reglamento General de Protección de Datos (RGPD) y aplicándolo cuando se trata de datos en los que se pueden identificar sujetos.

6 out of 6 points

rea 3: Enseanza y aprendizaje

Score for this Section: 24/24

Valoro con atención cómo, cuándo y por qué usar tecnologías digitales en el aula con mis estudiantes, para garantizar que aporten valor añadido

Your answer

✔ Utilizo herramientas digitales para implementar metodologías docentes innovadoras y compartirlas con mis redes, para que también puedan beneficiarse

No olvide reflexionar continuamente sobre la idoneidad de sus estrategias de enseñanza. Sea flexible, continúe ajustando su repertorio de estrategias digitales y pedagógicas y adapte su enseñanza a las necesidades de sus estudiantes. En su nivel de competencia, puede probar y validar métodos de enseñanza innovadores y compartirlos en su red profesional, para que ellos puedan también beneficiarse.

6 out of 6 points

Superviso las actividades e interacciones de mis estudiantes en los entornos colaborativos en línea que utilizamos

Your answer

✔ Redirijo la actividad en línea de los estudiantes cuando veo que no funciona o preveo problemas

Propone actividades en línea a sus estudiantes y sigue de cerca sus interacciones. Cuando ve que la actividad no funciona bien o sus interacciones no son las que esperaba, puede redirigir la actividad para aprovechar al máximo la tarea propuesta.

6 out of 6 points

Cuando mis estudiantes trabajan en grupo, utilizan tecnologías digitales para adquirir y plasmar los conocimientos

Your answer

✔ Diseño actividades curriculares que requieren el uso de tecnologías digitales para mejorar el aprendizaje colaborativo y la creación conjunta y el intercambio de conocimientos

Si puede incluir el uso de tecnologías digitales con los estudiantes a nivel curricular, asegúrese de que los estudiantes de su institución tengan la oportunidad de practicar actividades de aprendizaje que mejoran con el uso de las tecnologías. Esto les hará desarrollar sus propias habilidades en tecnología digital durante las horas extras.

6 out of 6 points

Utilizo tecnologías digitales para permitir a mis estudiantes planificar, documentar y monitorizar su propio proceso de aprendizaje Por ejemplo: autoevaluaciones, portafolios digitales para documentar y exponer, diarios/blogs en línea para reflexiones, etc.

Your answer ✔ Desarrollo aplicaciones o juegos digitales para involucrar a los estudiantes en su propio aprendizaje

De este modo, puede programar para aprovechar al máximo sus habilidades para integrar el uso de la tecnología en su docencia, de forma personalizada para sus estudiantes y áreas temáticas.

6 out of 6 points

rea 4: Evaluacin y retroalimentacin

Score for this Section: 18/18

Uso herramientas digitales de evaluación para monitorizar el progreso de los estudiantes

Your answer ✔ Desarrollo mis propias aplicaciones y herramientas digitales para seguir el progreso y/o realizar evaluaciones

De este modo, puede programar el mejor uso de sus competencias para integrar la tecnología en sus prácticas de evaluación, de forma personalizada para sus estudiantes y áreas temáticas.

6 out of 6 points

Analizo todos los datos disponibles para identificar de manera efectiva a los estudiantes que necesitan apoyo adicional Nota: "Datos" incluye: información personal, actividades de participación de los estudiantes, información sobre el rendimiento, calificaciones, asistencia e interacciones sociales en entornos (en línea); "Los estudiantes que necesitan apoyo adicional" son: estudiantes que están en riesgo de abandonar o tener un bajo rendimiento; estudiantes que tienen trastornos de aprendizaje o *necesidades educativas especiales; o estudiantes que carecen de habilidades transversales (p. ej., habilidades sociales, verbales o de estudio).

Your answer ✔ Animo a los estudiantes a no solo analizar los datos sobre su rendimiento, sino también a establecer sus propios objetivos de aprendizaje

Anime a los estudiantes a ser independientes, de modo que puedan establecer sus propios objetivos de aprendizaje y evaluar continuamente su propio rendimiento, buscando ayuda cuando sea necesario.

6 out of 6 points

Uso tecnologías digitales para proporcionar retroalimentación a los estudiantes

Your answer ✔ Desarrollo mis propias aplicaciones o herramientas digitales para proporcionar retroalimentación a los estudiantes

6 out of 6 points

De este modo, puede programar el mejor uso de sus habilidades para integrar la tecnología en sus prácticas de retroalimentación, de forma personalizada para sus estudiantes y áreas temáticas.

rea 5: Empoderamiento de los estudiantes

Score for this Section: 18/18

Cuando creo tareas digitales para los estudiantes, considero y abordo posibles dificultades prácticas o técnicas Por ejemplo: acceso igualitario a dispositivos y recursos digitales; problemas de interoperabilidad y conversión; falta de habilidades digitales

Your answer ✔ Selecciono y elijo herramientas que son accesibles e inclusivas, así como en formatos de código abierto para permitir una mayor personalización para mis estudiantes

6 out of 6 points

Usted conoce la importancia de las herramientas de código abierto y accesibles para personalizar la experiencia de los estudiantes.

Utilizo tecnologías digitales para ofrecer a los estudiantes opciones de aprendizaje personalizadas Por ejemplo: planteo diferentes tareas digitales a los estudiantes para abordar las necesidades de aprendizaje individuales, preferencias e intereses

Your answer ✔ Ayudo a los estudiantes a establecer objetivos y planificar las actividades que sienten que necesitan para mejorar su aprendizaje

6 out of 6 points

Comprende la necesidad de ayudar a los estudiantes a autoevaluar su progreso y poder establecer metas planificando sus propias actividades para el desarrollo.

Uso tecnologías digitales para que los estudiantes participen activamente en clase o en línea

Your answer

✔ Ayudo a los estudiantes no solo a crear, sino también a presentar y compartir el conocimiento que crean utilizando las licencias abiertas apropiadas

Además de ayudar a los estudiantes a estructurar, presentar y compartir el conocimiento que crean, les presenta el concepto de "bienes comunes" y les enseña a usar licencias abiertas para publicar su trabajo.

6 out of 6 points

rea 6: Desarrollo de la competencia digital de los estudiantes

Score for this Section: 30/30

Enseño a los estudiantes cómo evaluar la fiabilidad de la información

Your answer

✔ Debato con los estudiantes todo lo anterior y les enseño a no compartir información sesgada y engañosa

Ayuda a los estudiantes a identificar la distorsión de la información y la información errónea y sesgada. Esto hace que los estudiantes sean críticos con lo que leen y ven, por lo tanto, completamente capaces de evaluar la información.

6 out of 6 points

Configuro tareas que requieren que los estudiantes usen medios digitales para comunicarse y colaborar entre sí o con una audiencia externa

Your answer

✔ Animo a los estudiantes a mejorar sus habilidades de comunicación involucrando no solo a sus compañeros sino también a una audiencia externa como creadores conjuntos de conocimiento

Puede ayudar a los estudiantes a comunicarse más allá de su grupo inmediato de compañeros para aprovechar el conocimiento de una audiencia externa y, así, crear conocimiento de forma conjunta.

6 out of 6 points

Configuro tareas que requieran a los estudiantes crear contenido digital Por ejemplo: vídeos, audios, fotos, presentaciones digitales, blogs o wikis

Your answer ✔ Animo a los estudiantes a no solo a crear, sino también a compartir el conocimiento que generan utilizando las licencias abiertas adecuadas

Anima a los estudiantes a crear, adaptar y reutilizar contenido a la vez, compartiéndolo con un público más amplio a través de una licencia abierta.

6 out of 6 points

Enseño a los estudiantes a usar la tecnología digital de manera segura y responsable

Your answer ✔ Enseño a los estudiantes cómo detectar y evaluar la mala praxis en línea y las rutas para denunciarlo si se sienten personalmente ofendidos o atacados

Enseña a sus estudiantes a comportarse con confianza en línea y a detectar e identificar la mala praxis, así como a denunciarla si se sienten personalmente ofendidos.

6 out of 6 points

Animo a los estudiantes a utilizar las tecnologías digitales de manera creativa para resolver problemas concretos Por ejemplo, superar obstáculos o retos emergentes en el proceso de aprendizaje

Your answer ✔ Además de crear oportunidades para que los estudiantes utilicen sus habilidades digitales en la resolución de problemas, les dejo detectar estas oportunidades que surgen por sí mismos

Usted sabe la importancia de ayudar a los estudiantes a ser independientes y autocríticos, por lo tanto, también les permite detectar oportunidades para usar sus habilidades digitales en la resolución de problemas.

6 out of 6 points

rea 7: Educacin abierta (basada en el marco OpenEdu)

Score for this Section: 18/18

Sé cómo encontrar y utilizar licencias abiertas en recursos educativos

Your answer ✔ No solo uso licencias en abierto y comparto los recursos que creo, sino que también apoyo a mi institución en la implementación de REA como una práctica de educación en abierto

No solo acepta los REA en su propia práctica, sino que también contribuye a una política institucional que cumpla con los REA. Ayuda a sus compañeros a comprender los principios y prácticas de REA.

6 out of 6 points

Adopto prácticas educativas abiertas en mi docencia para hacerla más inclusiva

Your answer

✔ Adopto diferentes prácticas educativas abiertas en mi docencia y apoyo a mi institución para abrir el acceso al contenido (REA) y cursos a todos los estudiantes

Puede animar a su institución a ser más abierta:
1. Abogando por una infraestructura adecuada para los profesores que tienen como objetivo ofrecer REA, MOOCS y cursos en línea gratuitos y abiertos;
2. Asegurarse de crear y promover contenido y cursos variados, como en idiomas menos utilizados, así como para diferentes grupos de usuarios;
3. Crear programas de estudios para sus cursos que se puedan completar de forma modular, lo que permite una mayor flexibilidad (por ejemplo, micro-credenciales abiertos);
4. Alinear los programas de sus cursos con los de otras instituciones que ofrecen programas similares para permitir diferentes itinerarios de aprendizaje para estudiantes y la posibilidad de movilidad virtual.

6 out of 6 points

Publico mi investigación en revistas científicas abiertas, así como mis datos de investigación siempre que sea posible

Your answer

✔ Apoyo a mi institución en el diseño y cumplimiento de políticas que promuevan y/o recompensen a los profesores que adoptan la Ciencia Abierta y las prácticas de investigación abierta

Actúa como embajador de las prácticas de "Investigación Abierta" en su institución. Siendo ejemplo para sus compañeros, les ofrece orientación y comparte su experiencia con los equipos institucionales competentes para la toma de decisiones. Su objetivo es prestar el apoyo adecuado para las prácticas de "Investigación Abierta", así como la publicación en revistas de acceso abierto y la publicación de datos en abierto.

6 out of 6 points

Es usted…

Your answer ✔ Prefiero no decirlo

0 out of 0 points

¿Cuál es su edad?

Your answer ✔ Prefiero no decirlo

0 out of 0 points

¿En cuál de las siguientes ramas desempeña su docencia?

Your answer ✔ Otros (especificar)

0 out of 0 points

Especificar:

Your answer ✔ no answer given

0 out of 0 points

Incluyendo este curso académico, ¿Cuántos años lleva trabajando como profesor?

Your answer ✔ Más de 20

0 out of 0 points

¿En qué porcentaje de sus clases ha utilizado tecnología o herramientas digitales en los últimos tres meses?

Your answer ✔ Prefiero no responder

0 out of 0 points

¿Cuánto tiempo lleva utilizando tecnología en sus clases?

Your answer ✔ Prefiero no responder

0 out of 0 points

¿Qué herramientas digitales ha utilizado usted o sus estudiantes para enseñar y aprender? Puede seleccionar varias opciones

Your answer	✔ Otro ✖ Not all correct answers selected	0 out of 0 points

Especificar:

Your answer		0 out of 0 points

¿Qué edades tienen los estudiantes a los que enseña?

Your answer	✔ Mayores de 45 ✖ Not all correct answers selected	0 out of 0 points

¿Cuál es el perfil principal de sus estudiantes? Puede seleccionar varias opciones

Your answer	✔ Formación permanente (formación a lo largo de la vida) ✖ Not all correct answers selected	0 out of 0 points

¿Cómo evalúa su competencia digital docente como profesor ahora, después de responder al cuestionario? Asigne un nivel de competencia de A1 a C2, donde A1 es el más bajo y C2 el más alto Probablemente soy:

Your answer	✔ C2: Pionero	0 out of 0 points

Contact	jrc-digcompedu@ec.europa.eu
Useful links	DigCompEdu OpenEdu
Background Documents	Glosario Guia Educación Abierta
Contribution ID	98023d12-ee4c-43ce-aba0-5699b220629a
Completed at	19/08/2024 10:33:02
Completion time	-