El Zen i l'Art de la Intel·ligència Artificial

Enric Plaza

Butlletí de l'ACIA, 27 (Estiu 2002)

La intel·ligència artificial (IA) és l'enemic. Ho hem vist a les pel·lícules. A *Terminator* una IA desencadena la guerra final quan pren auto-consciència i sap que els humans volen destruir-la. També a *Matrix*, la IA redueix la humanitat al rol de bateries energètiques després d'una guerra que destrueix l'ecologia planetària. Aquestes pel·lícules no només busquen un enemic fàcil per fer guions efectistes, sinó que expressen una por davant un futur canviant: al segle passat les pel·lícules on l'enemic era el *científic boig* que volia governar/destruir el món expressaven la por per l'aparició de la Bomba A i la destrucció mútua assegurada (MAD). Podem pensar que aquestes pel·lícules simplement cometen un error en fer que l'enemic fos un científic boig en lloc en lloc de la política global de les superpotències amb armes nuclears. Tanmateix, aquestes pel·lícules també expressaven un rebuig de molta gent al desenvolupament de l'energia atòmica per la ciència i la tecnologia. Molts intel·lectuals i científics varen exigir excuses públiques als científics que van col·laborar en el Projecte Manhattan o van aconsellar als U.S.A. de desenvolupar l'arma atòmica en la lluita contra el nazisme a la 2ª Guerra Mundial. Semblantment, les pel·lícules on l'enemic és la IA també reflecteixen un rebuig de molta gent envers la IA. Aquest rebuig també s'expressa manta vegades en taules rodones i en les publicacions periòdiques quan després d'un article sobre la IA sempre hi ha cartes al director sobre les afirmacions que es fan sobre la IA. Tanmateix, la IA no ha posat el món al caire de la destrucció mútua assegurada. Per què hi ha doncs aquest rebuig a la IA?

Les revolucions científiques han erosionat el concepte que l'home té de si mateix. Encara avui diversos grups s'oposen a la teoria de l'evolució des de posicions religioses, filosòfiques i polítiques. La raó de fons és que considerar que l'home *només* és un animal molesta l'auto-imatge de molta gent. De la mateixa manera, la visió que la nostra ment (o la nostra *ànima*) sigui *només* un programa informàtic vulnera l'auto-imatge de l'home del segle XXI, Aquesta oposició proclama que "hi ha d'haver alguna cosa més". Un exemple històric d'aquesta oposició és el *vitalisme*, que fa cent anys proclamava que *la vida* no es podia explicar *només* amb mecanismes físics i químics. Avui aquesta posició ha estat superada en assolir la biologia molecular un model força clar del mecanisme dels éssers vius.

Malgrat tot, molta gent continua sentint enuig davant les manifestacions que descriuen els éssers vius només com mecanismes (en aquest cas mecanismes bioquímics). Els biòlegs opten usualment per expressar les nocions de la biologia de manera que no es faci palès aquest *mecanicisme* —per bé que no sempre és possible, com en les discussions sobre el "determinisme" dels gens en la fabricació del cos humà i les seves característiques. Irònicament, la biologia genòmica ha hagut d'importar el concepte informàtic/matemàtic de *codi* en els models sobre el "codi genètic": potser els mecanismes fisico-químics no són suficients per explicar els sistemes vius, com deien els vitalistes, però en un altre sentit? De fet, hi ha dues escoles dins la biologia moderna: la primera accepta la necessitat de usar la noció de *codi genètic* i la segona ho considera simplement una "metàfora", una forma de parlar, de la qual podrà prescindir-s'en quan s'acabin d'entendre els mecanismes genòmics.

En lloc de preguntar-se què és la IA o la intel·ligència aquest article té com objectiu preguntar-se per què l'estudi de la intel·ligència causa l'enuig, àdhuc el rebuig—preguntar-se quins problemes hi ha en la percepció de la IA (i veurem que també en la percepció de la "intel·ligència") per part de la gent. La manera de fer-ho serà examinar les diferents argumentacions que habitualment ens trobem en parlar de la IA. El propòsit no és per decidir quina escola dins la IA, la ciència cognitiva, o la filosofia de la ment, té la raó —sinó per elucidar els problemes que tenim tots plegats per pensar, per discutir, de la intel·ligència. Espero mostrar, finalment, que simplement deixar de parlar de IA (i

anomenar-la "informàtica avançada") no resol el problema, car la qüestió de fons és el substantiu *intel·ligència* i no el qualificatiu *artificial*.

Intel·ligències en plural

Comencem preguntant-nos per què ha de ser un problema que una màquina (un ordinador) tingui intel·ligència i no que tingui memòria. Sembla que el fet que una màquina tingui memòria artificial no ens posa cap problema, malgrat que la memòria d'una persona i d'un ordinador són coses molt diferents. Potser és pel fet que considerem normal que els animals "inferiors" a l'home tinguin memòria mentre que considerem que la intel·ligència de l'home és molt diferent de la dels animals: la nostra és l'autèntica i parlem només metafòricament de la dels animals. La paraula intel·ligència va lligada a d'altres com ment, consciència, racionalitat, pensament, etc., que la fan molt carregada de significat. A les cultures occidentals, des dels grecs els éssers humans es defineixen per contraposició a la resta d'animals com l'"animal racional". Per tant tenim a Occident una dicotomia fundacional en la imatge del que és l'home. Molta gent s'enutja o rebutja el concepte mateix d'intel·ligència artificial perquè viola aquesta dicotomia fundacional. A més, Occident té la tradició cristiana on el concepte d'*ànima* (molt semblant al de *ment*) també defineix l'home en contraposició a les altres "criatures" del món—que no en tenen.

El primer punt a l'ordre del dia és doncs sortir-nos de la dicotomia que reifica la intel·ligència com una cosa que només hi és en un pol i la seva absència que defineix el pol oposat. La teoria de l'evolució ja va destruir aquesta dicotomia en versions més antigues demostrant que de fet hi ha un continu entre (la resta de) els animals i nosaltres. Podem fer el mateix pas avui amb la intel·ligència i considerar que hi ha diferents tipus o nivells d'intel·ligència en els animals, inclosos els humans. Això equival a afirmar que existeixen *intel·ligències* en lloc de *La Intel·ligència* i parlar de la intel·ligència humana dins aquest continu. D'altra banda, els coneixements d'avui dia sobre la intel·ligència animal farien molt difícil mantenir la posició contrària: que els animals no tenen cap mena d'intel·ligència.

Tanmateix, no cal creure que això no ens elimina problemes: de fet ens il·lustra els problemes de fons. Els problemes que planteja l'estudi de la intel·ligència animal són, per exemple, si certs animals tenen llenguatge (o proto-llenguatge) o no; si es pot dir o no que aquests animals "entenen" el llenguatge que usen, etc. Des del punt de vista filosòfic ens podem plantejar els problema de si (o quan) podem dir que els animals "pensen" o tenen una ment. De fet, si ens fixem, aquests problemes són exactament les que apareixen quan discutim de la IA: en quines condicions podem dir que una màquina pensa, entén, o que té una ment? Aquesta coincidència evidentment no és fortuïta: el problema de fons no és la IA, el problema de fons són els conceptes "ment", "entendre" i "intel·ligència" tal i com els usem a Occident.

La ment del filòsof

Quan hi ha debats sobre la IA en els medis de difusió o es celebren taules rodones sobre el tema sempre sorgeixen les mateixes qüestions sobre si les màquines són intel·ligents, si pensen, o no. La discussió que segueix i els arguments que s'usen són de natura filosòfica, amb l'afegitó que els tertulians no són filòsofs. Malauradament, els filòsofs tampoc tenen la solució a aquests problemes car diferents escoles entenen els conceptes "ment", "entendre" i "intel·ligència" de formes molt variades. Tampoc no és l'objectiu d'aquest article advocar per una o altre proposta de la filosofia de la ment, però és important entendre que si discutim de certs temes no podem defugir el nivell filosòfic.

Per tal com la filosofia es centra en l'"home" com a objecte d'estudi hi ha poca feina feta respecte de si els animals no humans poden tenir ment i entendre una llengua. Per aquest motiu, la filosofia s'ha centrat en el debat de la ment i el cervell (humans). Podem resumir, seguint G. Rey (1997) en tres les posicions filosòfiques actuals respecte la filosofia de la ment: el reduccionisme, el dualisme i l'eliminativisme. Des de Descartes el *dualisme* ha

impregnat la "filosofia popular" a Occident, i és implícitament present en les discussions habituals sobre IA. "L'ànima racional no es pot extreure de cap manera del poder de la matèria, ans ha d'haver estat creat exprés". El dualisme de Descartes, que propugnava dues "substàncies" diferents pels fenòmens físics i pels mental, no és gaire popular a la filosofia contemporània. Les opcions reduccionista i eliminativista són les dues materialistes (o "fisicalistes") en el sentit que reconeixen la unitat dels fenòmens. Així, el reduccionisme diu que els fenòmens mentals són explicables (reductibles) a l'activitat cerebral, mentre que l'eliminativisme (a hores d'ara molt en voga) declara que els nostres conceptes sobre allò "mental" són simple folklore i que desapareixeran quan tinguem una comprensió científica del cervell.

El debat filosòfic en aquests termes resulta decebedorament poc il·luminador dels problemes que són realment interessants. La idea de reduccionisme no ens ajuda gaire: potser la vida "es redueix" a la física, i la ment al cervell però sembla que seguirem fent models biològics i psicològics dels fenòmens. També els software "es redueix" al hardware —però és impossible parlar amb sentit de certes propietats amb el vocabulari pertinent al hardware. L'eliminativisme ve a dir que no tan sols té sentit parlar de software, per seguir l'analogia, quan podem fer un bon model del hardware. En informàtica es va començar a nivell de hardware, i nous nivells de descripció conceptual van haver d'ésser desenvolupats (i implementats), i són aquests nivells els que genèricament anomenem software. Va caldre inventar un mot per descriure els processos i fenòmens complexos que emergien d'una estructura tan aparentment senzilla com un "computador de propòsit general". Certament, la filosofia no s'ocupa d'aquests nivells de descripció, d'això s'ocupen les ciències com la psicologia —però en els debats sempre retornem als conceptes filosòfics!

La qüestió fonamental que sembla definir la "ment" en debat filosòfic actual és la *intencionalitat*. Aquest mot és poc intuïtiu perquè no té a veure directament amb les intencions o objectius d'algú. La intencionalitat és la capacitat per la qual certs continguts (les "actituds proposicionals") són *sobre altres coses*, és a dir tenen significat o contingut semàntic. Fixem-nos que aquest plantejament fa entrar com component essencial la idea de llenguatge (humà) i la capacitat de comprensió ("entendre el significat"). Sovint aquest esquema filosòfic afirma que els programes informàtics són *sintàctics* només, en ser incapaços de "entendre" el significat. Tanmateix, el problema que es volia resoldre (caracteritzar la ment) fa entrar tots els problemes relacionats: llenguatge i comprensió caracteritzats únicament com aquells típicament humans. Aquesta circularitat ve a dir que tens una ment si ets un humà, i si no, no. Que en fem de Washoe, el ximpanzé bonovo que va aprendre a comunicar-se amb l'ASL (American Sign Language)? Com podriem esbrinar si ell té intencionalitat? I un mico o una rata? Tal com es defineixen intencionalitat, semàntica i comprensió, poc es pot dir aplicat a aquests animals. Si no tenen ment són mecanismes sintàctics com un program informàtic? Potser una rata és "només" sintàctica i Washoe té "una mena d'intencionalitat". En aquests supòsit, quines menes d'intencionalitat poden haver? El problema aquí torna a ser la qüestió encara debatuda pels biòlegs de si els animals "poden tenir o no" llenguatge—i es defineix llenguatge *només* com el llenguatge humà.

Fins ara hem examinat la problemàtica de la ment i hem dit poc de la IA. Certament, hom pot acceptar una visió materialista de la ment i la intel·ligència negant la pertinència d'usar el mot intel·ligència quan es parlar de la informàtica, els ordinadors, la IA, i tota la patuleia. Des de la filosofia, la IA ve qualificada dins la tendència anomenada funcionalisme. El *funcionalisme* diu simplement que certes coses són el que són no per la substància de la qual són fetes, sinó per la funció o rol que desenvolupen en un sistema més ampli. Per exemple, que alguna cosa sigui diners no depèn de si està feta d'or, de paper, o de senyals electròniques en un ordinador, sinó del rol que juga en l'intercanvi de bens dins un mercat. La IA adopta una posició funcionalista en afirmar que les substàncies diferents entre cervell i ordinador no importen mentre es demostri que poden jugar el mateix rol: jugar a escacs, pilotar avions, etc. Molts dels atacs contra la IA des del punt de vista filosòfic són de fet atacs contra el funcionalisme. Per tal com l'objectiu d'aquest article no és debatre quina posició filosòfica és la correcta, sinó d'esbrinar els problemes conceptuals que ens trobem en intentar pensar i explicar la IA, deixem ara la discussió sobre les posicions filosòfiques. A partir d'ara examinarem els arguments més habituals amb els quals es discuteix si la IA

pot ser o pot tenir sentit. Aquests arguments s'examinaran des del punt de vista que d'*intel·ligència* i *ment* són problemes oberts més que coses clares que s'utilitzen per explicar què és, o no és, o no pot ser, la intel·ligència artificial.

Intel·ligència simulada

Un argument sovint esgrimit és que un sistema de IA no és realment intel·ligent sinó que "simula la intel·ligència". De fet, l'enunciat comet un error en confondre les nocions de simulació i model (de l'àmbit teòric) amb un programa informàtic que realitza accions en el món. Si jugo a l'ordinador amb un programa d'escacs, aquest programa no *simula* jugar, està *efectivament* jugant amb mi. De fet, una persona juga als escacs amb peces físiques o visualitzades en una pantalla d'ordinador. Un pilot automàtic dirigeix *efectivament* un avió quan el pilot li cedeix el control, no és un simulador de vol, ni "simula" un pilot: *és* el pilot. És veritat que podem fer programes informàtics que són models d'activitats o fenòmens: aquests programes són simulacions perquè són models. Un programa deixa de ser una simulació quan és *causalment* unit al món extern: el programa mou la torre (sigui la peça virtual o material) i em fa escac, el programa gira l'avió a l'esquerra, són accions en el món. Per tant, un programa pot simular o pot actuar, depenen del seu lligam causal amb l'entorn. Podem discutir si un pilot automàtic o un ocell tenen ment, però *volen*, no simulen.

Aquesta confusió la repeteix Searle una i altre vegada des de la seva "habitació xinesa". En una entrevista, preguntat per uns hipotètics robots que fossin capaços de parlar i entendre's amb els humans, Searle diu que aquests robots serien "simulacions informàtiques de patrons de conducta dels éssers humans". Que un filòsof pugui confondre un simulador de vol amb un pilot automàtic em deixa esbalaït. La posició de Searle fa que la recerca en vida artificial i intel·ligència d'eixam (*swarm intelligence*) siguin tant inútils com la IA. Per posicions com les de Searle la vida artificial és simulació i no vida, com la IA és simulació i no intel·ligència independentment de tot. El problema, però, és que això no és independent: els virus informàtics que pul·lulen per la no són simulats perquè efectivament causen molts danys. Els experiments en vida artificial són simulacions perquè, molt assenyadament, s'isolen del món real (Internet i Windows en aquest cas). La prova d'això és que consideraríem un acte criminal que algun investigador de vida artificial llencés al món poblacions de "éssers" que fossin capaços de reproducció i evolució genètica explotant els recursos d'Internet i Windows. El projecte Tierra de Tom Ray sobre evolució definia un "codi màquina" a partir del qual evolucionaven unes poblacions d'éssers composats per instruccions que permeties llur auto-reproducció i mutació; doncs aquest "codi màquina" era diferent del codi màquina real precisament per evitar una contaminació del món real des de la plataforma experimental—precaució assenyada que tot investigador sobre la vida (natural o artificial) ha de prendre imprescindiblement. Per cert, avui dia la comunitat de recerca en biologia encara no té un consens sobre la definició de vida, ni les característiques que defineixen un ésser com *ésser vivent*. Concretament, no se sap si els virus, els retrovirus i els prions haurien d'ésser inclosos dins la categoria de éssers vivents (o de la "matèria inerta", el problema és que són massa actius per considerar-los *inerts*). Semblaria aquest un problema filosòfic, però sortosament per la biologia contemporània ningú sembla preocupat per debatre aquest tema.

El mecanisme del programa

Aquesta confusió conceptual entre model de simulació i sistema que actua en el món real està molt relacionada amb un altre aspecte que ara destacarem: els programes es consideren entitats abstractes i no com a *mecanismes*. La percepció de la gent del carrer és que els programes (i els ordinadors) són ens *lògics*—amb totes les connotacions que té aquest mot. Els acadèmics del software també fan l'èmfasi en l'aspecte més formal, cosa important quan es dissenya un algorisme. Per tal d'argumentar la visió dels programes informàtics com mecanismes cal distingir tres aspectes: l'algorisme, la implementació i el procés. Per *algorisme* entendrem l'especificació d'un procés o activitat en algun llenguatge, per *implementació* entendrem la codificació d'un procés o activitat, i per *procés* entendrem una execució de la implementació en un ordinador. Quan hom diu que un programa és una cosa

únicament sintàctica pensa en l'algorisme o la implementació. Un procés no pot caracteritzar-se només de sintàctic perquè té una part *pragmàtica*: realitza les accions en el món real (juga als escacs, pilota un avió). Queda sempre el debat obert de si se li pot assignar un contingut semàntic: la informàtica ortodoxa diu que els programes tenen una semàntica operacional mentre que la semàntica en el sentit intencional de la filosofia és l'etern debat sobre la IA.

Centrem-nos en el *programa com a procés* i veurem que és essencialment un *mecanisme* (o bé, com es diu sovint, un *sistema*). Fixem-nos que qualsevol programa software que s'executa en un computador d'ús general pot traduir-se a un "mecanisme" hardware compost per circuits que són equivalents (en el sentit funcional) al programa original. El software es desenvolupa simplement perquè és una manera més ràpida, barata i flexible de construir mecanismes, però conceptualment és equivalent a construir un mecanisme específic. De fet, un programa en el sentit "sintàctic" és una especificació d'un mecanisme, mentre que un ordinador (que els primers creadors anomenares *computador d'ús general* per distingir-lo dels computadors especialitzats amb la circuiteria per certes tasques) és una màquina universal que pot *realitzar* el mecanisme de qualsevol especificació sintàcticament vàlida. Certament, els "sistemes" que es desenvolupen tenen no només una part software ans també una part hardware: sensors, actuadors, monitorització de processos, qualsevol perifèric que es pugui construir. El sistema total és el mecanisme del que parlem: el pilot automàtic és el sistema complet de sensors i actuadors a més del "programa de presa de decisions"; el pilot automàtic és exactament el mateix si la presa de decisions la fa un programa software o un mòdul de circuits: ambdós són implementacions del mateix mecanisme. De fet, si pensem en el pilot automàtic compost per mòduls totalment fets de circuits hardware tindrem més clara la intuïció que es tracta d'un mecanisme, d'una *màquina*. Fixem-nos, per últim, que quan diem que són iguals un programa software i un circuit adoptem una visió funcional: ambdós fan el mateix malgrat són fets d'estofes diferents.

Aquí cal deturar-se en el concepte de màquina i afegir que l'usem en un sentit abstracte, com un sistema material complex (on els components poden ser mecànics, electrònics, etc). Des del punt de vista filosòfic, la primera crítica al dualisme de Descartes que reivindica la visió materialista (fisicalista) de l'home és La Mettrie en el seu assaig *L'Homme Machine.*En essència, La Mettrie diu que la ment no és una entitat separada de la matèria i que per tant l'home és una màquina (o, si voleu, un "sistema material"). Cal remarcar que aquest concepte és mol important en el neguit que la IA desperta en la gent (siguin filòsofs o no) degut a que modifica la visió que l'home, afirmant que nosaltres "som només una màquina". El neguit porta al rebuig, i a cercar argumentacions per les quals nosaltres no som (no podem ser) *iguals* a màquines —o bé, filant més prim, perquè nosaltres, entitats materials biològiques no som (no podem ser) *iguals* a sistemes mecànics/electrònics. La posició de la ciència, clarament, no pot ser aquesta car no pot prendre cap a priori sobre qüestions empíriques i aquesta n'és una.

La qüestió substancial

Un argument contra la visió que nosaltres, com a mecanismes, som equivalents màquines electròniques és qualificar aquesta visió de funcionalista (que ho és) i afirmar que la substància és la que crea la diferència. En altres paraules, els éssers vius fets de carboni i els sistemes electrònics fets de silici no podran mai tenir les mateixes propietats car les substàncies no ho permetran. Si això és cert o no, és una qüestió empírica, car no es pot saber a priori si una certa organització de la matèria basada en silici pot mostrar certes propietats. Des del punt de vista pràctic, és versemblant poder construir en el futur robots amb la intel·ligència d'un ratolí, per exemple. L'argument de la substància aquí segurament afirmaria que el ratolí robot i el ratolí animal "no són el mateix" perquè el robot fa les tasques del ratolí sense ésser un ratolí: no sap el que és "ésser un ratolí" car no té necessitat d'alimentar-se igual, ni de reproduir-se. La suposada igualtat és només *funcional* i això és el que l'argument nega, aquesta és la qüestió essencial i no la substància i les seves propietats. Searle diria que el robot ratolí simula un ratolí car no *és* un ratolí. Resumint, la discussió al final es centra en què volem dir quan parlem d'igualtat entre dues coses, si és una igualat funcional o és una igualtat fenomenològica. Finalment, sembla una discussió

ociosa, perquè la IA sempre adopta una posició funcional i mai ha afirmat la igualtat fenomenològica (mai ningú no ha dit que un ratolí robot *és* un ratolí). Recordem la idea que diferents animals tenen diferents nivells/tipus d'intel·ligència: aquest és el punt de vista que interessa a la IA. Ara suposem que un ratolí i un rat penat tinguessin el mateix tipus o nivell d'intel·ligència: ningú no afirmaria que un rat penat *és* un ratolí fenomenològicament (l'experiència de l'entorn és diferent amb sistemes de percepció diferents, per començar). Tanmateix, sí que es podria afirmar, dins una teoria de la intel·ligència animal, que rat penat i ratolí tenen *la mateixa* intel·ligència—mentre que la posició dels filòsofs contraris [*contrarians*] no els ho permetria sense caure en contradicció.

Encara més, la qüestió de la substància no tanca el debat sobre la intel·ligència artificial o mecànica, car el futur pot desenvolupar computadors d'ús universal en d'altres substàncies. A *Les noves ments de l'emperador* s'afirma la hipòtesi que el pensament creatiu humà té una base en la mecànica quàntica. No es diu quina part del mecanisme neural es veuria afectat quànticament ni quin seria l'efecte, però suposem que sigui una hipòtesi versemblant que es demostrés certa en el futur. Això faria impossible la IA o simplement assenyalaria que caldria construir ordinadors que incorporen un cert efecte quàntic de la manera apropiada? La IA no està lligada, conceptualment, als ordinadors actuals: els ordinadors són un instrument de la IA com el telescopi és un instrument de l'astronomia. De fet, la història de la informàtica es pot interpretar des de la història de la IA quan encara no existia i es deia cibernètica. Els pioners de la cibernètica varen iniciar el *programa de recerca* (en el sentit de Lakatos) de la intel·ligència mecànica i com a conseqüència d'aquest programa es varen crear els conceptes fonamentals de la informàtica: els circuits lògics venen de McCulloch (*A logical calculus of the idea immanent in neural nets*) i l'arquitectura bàsica de l'ordinador és la de Von Neumann. La IA redefineix el programa de recerca de la cibernètica amb la proposta de centrar-se en el software com a millor manera (en una època on els ordinadors comencen a ésser presents a l'universitat) de dissenyar mecanismes per tasques intel·ligents.

El propòsit de tot plegat

Una qüestió relacionada és la del *purposeful behavior*, és a dir la qüestió de si els mecanismes poden tenir propòsits, conductes guiades per objectius. La visió tradicional és que només els homes en tenen i totes les demés coses (incloses els animals) no en tenen. Tornem a tenir una dicotomia estricte, i la manera de sortir-s'en és veure que pot haver un continu des de conductes amb propòsit senzilles fins a les més complexes que es donen en els humans. Avui sabem que els animals tenen propòsits (*purposeful behaviors*), per bé que la biologia els anomena púdicament "tropismes" i l'etologia parla de "conductes".

La visió clàssica recrea la dicotomia distingint entre propòsits humans (suposadament propis, lliures i conscients) i els dels animals (uns propòsits derivats de l'evolució i per tant no-propis dels individus). Aquesta dicotomia sembla dividir els propòsits entre els exogens i els endogens (que serien els pròpiament humans). Tanmateix, els objectius dels humans venen molt determinats per la cultura i la societat—és més, els "objectius bàsics" dels humans són biològics i per tant determinats per la evolució. Aquí retrobem la dicotomia clàssica entre natura i cultura (*nature* vs. *nurture*) car un nadó té unes capacitats i un "temperament" que venen determinades genèticament mentre que altres capacitats i la personalitat adulta venen determinades per la cultura. L'objectiu d'estudiar a la universitat i trobar una nova feina és un objectiu endogen (ho faig perquè vull) o exogen (és una exigència social per certes tasques). També hi ha objectius més íntims, és clar: per exemple tenir un fill .... és clar que això és un objectiu biològic. Com resoldre la qüestió? Cal considerar la dicotomia falsa i reconèixer que els objectius dels humans tenen sempre els dos aspectes. La psicologia ha estudiat els "projectes de vida" que una persona es fa, amb uns objectius genèrics i altre més específics per assolir els genèrics. Si bé l'assumpció de certs objectius com més importants és individual (obtenir un cert status social abans que tenir fills) els objectius genèrics com a tals són sempre d'orígen biològic (tenir fills) o cultural (obtenir un status, un reconeixement social).

Si retornem al concepte de *purposeful behavior* veurem que des del punt de vista de la IA s'ha anomenat conducta guiada per objectius (*goal-driven behavior*). Newell va afirmar que

els investigadors en IA descrivien els sistemes desenvolupats a dos nivells: al *nivell de coneixements* i al *nivell de mecanisme*. El nivell de coneixements descriu els sistemes de IA *com si* tinguessin objectius i expliquen les accions que realitzen amb el principi de racionalitat: el sistema usa les accions que sap que poden assolir els objectius. Així, el sistema es descriu *com si* fos un subjecte que té intencions i realitza les accions que racionalment "sap" que assoliran "els seus" objectius. Naturalment, els objectius són externs (donats pel dissenyador) i les decisions són "mecàniques" (és a dir, són codificades en el mecanisme del programa). Per tant, les descripcions de sistemes de IA es fan a dos nivells: a un nivell intencional i a un nivell de mecanisme. Això ha portat a moltes discussions en IA de tipus *normatiu*: la discussió de si és correcte fer-ho així o no. Si deixem de banda aquest debat normatiu, ens podem demanar perquè la gent fa aquestes descripcions quan descriu sistemes de IA. En Newell ho deixa clar quan explica que és la manera més senzilla d'explicar coses complexes (el mecanisme). Però això no és privatiu dels sistemes de IA, moltes explicacions de sistemes de software també els descriuen com subjectes (agents) que realitzen accions per tal d'assolir objectius. No només això, les descripcions intencionals s'usen habitualment en les nostres descripcions de les conductes dels animals: "la gasela sap que darrera la roca i ha un lleó". Tanmateix, tot això és incorrecte segons la filosofia heretada on només els humans tenen ment: ni un sistema d'IA ni una gasela "saben" res ni "tenen intencions".

Recordem ara les opcions reduccionista i eliminativista de la filosofia de la ment contemporània. Segons l'opció que escollim, la relació entre els nivells intencional i de mecanisme és diferent. El reduccionista ens dirà que hi ha un descripció intencional però que es pot "reduir" (en el sentit d'explicar) a un nivell on s'explica el mecanisme del sistema material. En la gasela això es pot fer (en principi, per bé que no actualment) i en el cas de sistemes informàtics això es pot fer sempre. L'eliminativista ens diu que descrivim així la gasela (com també les persones) perquè no coneixem el mecanisme subjacent, i que si el coneguéssim no caldria aquesta descripció intencional. Respecte els sistemes informàtics, aquesta visió simplement no entén perquè s'usa el nivell intencional (donat que coneixem perfectament el mecanisme: nosaltres l'hem dissenyat). Probablement, l'eliminativista diria que aquesta descripció intencional és *misguiding* o incorrecta. Malauradament, l'eliminativisme no pot explicar perquè necessitem fer les descripcions intencionals (funcionals) quan en informàtica *tot el sistema material es coneix perfectament*. No estarem perdent quelcom de vista? No són potser les descripcions de *purposeful behaviors* irreductibles, i cal per tant la descripció del mecanisme però també la descripció del *propòsit* del mecanisme?

La diferència entre reduccionistes i eliminativistes sembla doncs centrar-se en l'estatut del que abans hem anomenat la descripció *funcionalista*. El reduccionisme accepta un nivell de descripció on s'usi el paradigma funcionalista, però sembla acceptar-lo de mal cor, tot afirmant que, de fet, es pot *reduir* a una explicació fonamental. L'eliminativista rebutja un nivell de descripció funcionalista, el considera espuri. El problema amb aquestes aproximacions filosòfiques és que prenen com posicions de principi qüestions que, de fet, no estan gens clares. El dualisme ment/matèria es pot resoldre amb les perspectives reduccionista i eliminativista? Considerem l'informàtica altre cop, on els sistemes materials constituïts de hardware es coneixem perfectament per disseny. Per què calia inventar-se la paraula software? En el seu moment, software es va encunyar per designar tot allò de què necessitàvem parlar i que no era, exactament, hardware. La pregunta queda aquí oberta: Què és, doncs, el software?

Com experiment mental, apliquem les categories del reduccionisme i eliminativisme a la distinció hardware/software. En primer lloc, aquesta distinció no és un dualisme, car es consideren, per dir-ho ras i curt, dos nivells de descripció pel mateix sistema material (ens referim aquí al *programa com a procés*, no al codi o a l'algorisme d'un programa com abstracció). Dir que un *programa com a procés* es pot *reduir* al hardware no sembla assenyat, ni sembla que ens doni nova llum sobre la relació entre software i hardware; més aviat sembla confondre. Si el reduccionisme només és una negació del dualisme, el mot *reducció* és sobrer, i s'hauria de simplement materialisme o fisicalisme. D'altra banda, proposar *eliminar* tota noció de programa o software per tal com ja tenim una teoria del hardware sembla forassenyat—però és el que sembla implicar la opció eliminativista. Si

l'eliminativisme, en canvi, acceptés que es pot parlar de software, com pot propugnar l'eliminació de tota noció de ment? El problema d'aquestes filosofies que heretem és que semblen no acceptar que l'important és que cal descriure la realitat en diferents *nivells d'organització*. Per què no accepten això? Per la raó que la noció d'organització és funcional, i aquestes nocions són les que de fet es volen rebutjar. El que tenen en comú ments (humanes i animals) amb els programes és que tenen *contingut*, on el contingut és una certa organització de la matèria que té *sentit* en l'interior del mecanisme de cada sistema. No té sentit eliminar (o reduir a hardware) les nocions de programa perquè el hardware d'un computador de propòsit general no és res concret: és un contenidor (universal) que admet qualsevol mecanisme expressable en el seu "codi".

Per últim, cal aclarir que el *purposeful behavior* és un fenomen emergent d'un mecanisme, En concret, no cal tenir un sistema amb *contingut* (un programa informàtic o un sistema nerviós central) per tal de tenir propòsits. Per exemple, pensem en les plantes que tenen una conducte que podem descriure així: "aquesta planta busca el sol". No importa que en biologia li diguin púdicament "tropisme" solar, la planta té efectivament un mecanisme (explicable en biologia) que genera la conducta que anomenem "buscar el sol". La discussió de si això cal dir-li o no "propòsit" no pot ser només una qüestió lingüística. La negació que animals (i plantes) tinguin "propòsits" té la mateixa raó que la negació que tinguin llenguatge o intel·ligència: aquestes prerrogatives només es poden adjudicar a l'home. De fet, aquesta negació és la conseqüència d'una altra: la negació que el propòsit (el llenguatge, la intel·ligència) siguin un *mecanisme*. El fet que hi ha mecanisme elimina la necessitat de parlar de propòsit, d'intenció? O és només una metàfora? Aquesta és la proposta de l'eliminativisme. Malauradament, les coses no són tan senzilles. La categoria de *purposeful behavior* va estar introduïda per la cibernètica a partir de la noció de retroacció (*feedback*). Conceptualment això va ser una revolució—car des del temps de la Grècia clàssica estava "prohibit" que els efectes guiessin de cap manera les causes. El que en la tradició es considerava un cercle viciós va esdevenir (en paraules d'Edgar Morin) un bucle virtuós: l'auto-regulació dels sistemes i el sorgiment de conductes *guiades* per un objectiu—i tot explicable per un mecanisme subjacent.

En arribats a aquest punt, hi ha dues alternatives en concebre la intel·ligència, la intencionalitat, i el llenguatge. La tradicional a Occident ha estat reservar aquests conceptes als humans i negar la seva aplicabilitat a d'altres "sistemes materials". La segona alternativa, l'anomenada perspectiva funcionalista, és que hi ha un continu entre sistemes materials senzills i els més complexos, i que hi ha diferents varietats de fenòmens que s'encuadren en el que comunament anomenem intel·ligència, llenguatge, i *purposeful behavior*. Aquesta perspectiva només pot ser funcionalista, car la equivalència que proposa es basa en la funcionalitat assolida per certs nivells de l'organització de la matèria. L'alternativa tradicional causa problemes no només a la IA sinó també a la biologia contemporània. En estudis sobre la "interacció" entre animals és científicament correcte parlar de *comunicació* però no de *llenguatge*, d'intercanvi de *senyals* però no de *comprensió* d'aquests senyals—tot perquè *llenguatge* i *comprensió* s'haurien de reservar (suposadament) a la ment humana. Però com pot haver comunicació sense comprensió? Com pot respondre algú (o alguna cosa) a una senyal sense entendre-la? La resposta de l'alternativa tradicional és que la ment humana és quelcom especial, i els altres sistemes materials són d'una altra mena: responen "per automatisme". Aquesta perspectiva obliga a parlar de la comunicació animal en el nivell de descripció de mecanisme: un animal rep certa senyal i cal descriure el mecanisme que efectua la resposta a partir d'aquest senyal. Donat que el nivell de descripció intencional està prohibit pels animals no humans, el mecanisme no "compren" la senyal ... simplement perquè no hi ha "ningú" que pugui entendre-la (no hi ha "subjecte"). De fet, la tradició vol mantenir que els humans "són alguna cosa més" a nivell mental, de la mateixa manera que els vitalistes del segle XIX mantenien que la vida era "alguna cosa més" que un sistema material.

La metàfora de la ment

El problema de tot plegat és que la perspectiva tradicional només és consistent amb la filosofia dualista i no amb la fisicalista (sigui reduccionista o eliminativista). Cal un continu entre la intel·ligència humana i la dels primats, entre la dels primats i els mamífers, i així

fins els sistemes vivents sense sistema nerviós central. El mateix continu és necessari per la qüestió de la comunicació, la de la comprensió, i la del llenguatge. Malgrat les reserves filosòfiques, en el dia a dia les ciències usen tant el nivell de descripció de mecanisme com el nivell intencional. Una objecció freqüent a les descripcions intencionals és que són "metafòriques". Dit altrament, quan diem que un animal avisa un altre d'un depredador, o quan diem que un pilot automàtic fa aterrar un avió, estem suposant un *subjecte* que no existeix al qual assignem *metafòricament* intencions que *realment* no tenen. Per tant, emprem el mot "metàfora" per significar que allò que es diu no és "real". L'eliminativisme, en essència, adopta aquesta posició fins i tot respecte la intel·ligència i la ment humana.

Deixem la qüestió del subjecte per la propera secció i concentrem-nos en la necessitat (o la sobreria) de parlar de "ments". Com hem vist, la filosofia de la ment gira en torn la noció *d'intencionalitat* (en el sentit de la capacitat d'entendre un llenguatge i que les expressions del llenguatge són sobre "alguna cosa"). Des del punt de vista de la ciència contemporània sembla necessari tenir un continu entre les capacitats cognitives dels animals incloent l'home. Malgrat que aquesta teoria de la cognició animal encara està per fer, el fet és que cal una continuïtat que no és consistent amb la visió dicotòmica on els homes poden comprendre i els altres éssers no. Des d'aquest punt vista hi ha una gamma de capacitats diferents a les quals podem anomenar *comprensió* i *llenguatge*. Conseqüentment, també diferents animals tindran diferents tipus de ment—de fet avui ja es parla de capacitats cognitives en els animals i àdhuc d'intel·ligència animal, però sembla que parlar de ment animal és problemàtic (com si fos pecat de panteisme: com si fos dir que els animals tenen *ànima*).

Hom pot pensar que els arguments basats en la intel·ligència animal no tenen cap força per aplicar-los a la intel·ligència artificial. Tanmateix, la qüestió en discussió aquí sorgeix tant en parlar de màquines com d'animals. Si no s'accepta que una màquina sigui intel·ligent (que "pensi", que tingui "ment") tampoc, en el marc filosòfic heretat, no s'accepta com correcte dir que un ratolí "pensi". Podriem considerar una posició "actualitzada" que admet que certs animals poden "pensar" (o tenir "ment") però no pas les màquines—una posició que acceptaria la intel·ligència natural però no l'artificial. Tanmateix, aquesta posició reformada també es basa en una dicotomia entre els que pensen i els que no, entre els que tenen ment i els que no; tornem a trobar la necessitat de marcar un límit: per exemple que els primats pensen i els altres animals no. O bé els mamífers sí que pensen però no pas els rèptils—perquè uns i amb quins criteris es defineix aquesta frontera? Aquesta posició "actualitzada" simplement actua sobre la l'autoimatge de l'home: allò que s'hi assembla més (primats, mamífers) és acceptat en el club de la ment, i la resta es rebutja. Hom pot argumentar que la investigació científica podrà empíricament delimitar aquesta frontera, després d'estudiar les habilitats cognitives d'aquestes espècies. Però precisament això és el que hem torbat abans que la biologia tenia problemes en definir (quan es podia dir que un animal "comprenia" un crit, un senyal—quan era un abús de llenguatge dir que un animal "comprenia"). El debat no és circumscrit a qüestions "tècniques" dins una ciència perquè estem tractant amb conceptes predefinits amb unes connotacions filosòfiques molt carregades: les conceptes heretats determinen quan es pot dir que un ésser "enten" o "pensa".

La perspectiva eliminativista és doncs atractiva: tot plegat són conceptes massa connotats i no són útils, caldria desenvolupar nous conceptes per l'estudi ... de la "ment". Si cal eliminar el concepte de la ment sobre què estem parlant? La sortida d'aquest atzucac és passar a parlar del *cervell* en lloc de parlar de la *ment*. És a dir, en els termes que hem introduït anteriorment, passar al nivell de descripció del mecanisme i deixar de banda el nivell de descripció funcional. De fet, la ciència cognitiva va fer un pas anàleg en denunciar com arcaics i inútils els conceptes heretats de la *folk psychology*. Probablement és un pas assenyat: una disciplina científica ha de provar de desenvolupar conceptes adequats al seu camp d'estudi. Tanmateix, aquest pas pot ser necessari però no suficient: no és clar que la descripció mecànica sigui suficient i que pugui evitar-se la descripció funcional. Si retornem a l'analogia informàtica, donat que coneixem el hardware totalment, quina necessitat hi ha d'introduir els conceptes que s'usen en la *software engineering*? Si coneguèssim totalment els mecanismes cerebral no seria sobrera la psicologia com a ciència? La idea que calen diferents nivells de descripció torna a fer-se palesa. Els models

desenvolupats per la psicologia i la informàtica del software no són sobrers: parlen del mateix sistema material però a un "nivell d'organització" diferent. Si són necessaris, i aquestes disciplines usen tant aviat models de mecanisme com models funcionals, ambdós són necessaris i correctes. Conseqüentment, cal revisar les definicions dels conceptes heretats de manera que siguin adequats a les necessitats dels models que es desenvolupen en aquestes ciències. Quan sigui necessari descriure situacions on algun ésser "compren" algun senyal, s'ha de poder acceptar l'argument, després de discutir-lo, en lloc de rebutjar-lo d'entrada.

El subjecte de la qüestió

La qüestió del *subjecte* és, no cal dir-ho, essencial en la filosofia occidental. En la visió que hem heretat, el subjecte es defineix circularment a partir de les nocions de persona i de comprensió (del llenguatge). Donat que només les persones són subjectes, altres sistemes materials, biològics o artificials, no poden "comprendre" i per tant tampoc no poden tenir cap "llenguatge". Perquè han d'anar lligats els conceptes de *subjecte* i *ment*? En principi no caldria, però el fet és que la definició de ment en la filosofia actual va lligada a la "capacitat de comprensió" (que s'anomena curiosament *intencionalitat*). Dit altrament, el subjecte és qui "compren", i per tant només un subjecte té "ment": ja hem lligat tots els conceptes en un entrellat indestriable. Tots aquests conceptes són auto-referents, però compliquen les qüestions obertes més que no ajuden a solucionar-les. Aquest vocabulari heretat forma un marc prescriptiu més que explicatiu. Prescriu que no podem parlar de "llenguatge" i de "comprensió" en els animals no humans, i de fet es pot veure en articles científics que els autors s'estimen més parlar només de "formes de comunicació" entre animals per tal d'evitar "problemes filosòfics". Per tant, la prescripció filosòfica està determinant el que la ciència "pot dir" i el que "no pot dir". És molt discutible que prohibir parlar de llenguatge i comprensió en qualsevol sistema material que no sigui l'home i permetre en canvi parlar de comunicació sigui la decisió correcte. És possible que sigui el decurs de la investigació científica (en IA, en biologia, en neurologia) on s'hagi de decidir quan cal parlar de llenguatge i de comprensió. Això vol dir que les nocions de *llenguatge* i *comprensió* (com *intel·ligència* i *ment*) deixen de ser nocions donades (clares i predefinides) i passen a ser conceptes problemàtics (i oberts a la discussió).

Molt probablement, el concepte de subjecte que hem heretat s'haurà de revisar, però no és aquest el propòsit d'aquest article debatre quines solucions hi ha als problemes filosòfics. De fet, els filòsofs contemporanis han debatut sobre la inadequació dels conceptes heretats com "subjecte" i "pensament", per bé que no semblen haver arribat ha un consens sobre la qüestió. Per exemple, Nietszche, a *La Voluntat de Poder*, nega l'existència de l'esperit (de la ment) com "una cosa que pensa"; el concepte de ment deriva d'una falsa introspecció que creu que el pensament és un acte i per tant cal una cosa que realitza aquest acte. Heidegger hi dedica dos assaigs: *Què significar pensar?* i *La fi de la filosofia i la tasca de pensar*. A més, concretament a *Lletra sobre l'Humanisme*, diu (p. 199): "Des d'aquesta perspectiva, "subjecte" i "objecte" són termes metafísics inadequats que han determinat, des de temps immemorial, la forma en què la "lògica" i la "gramàtica" occidentals han esbiaixat la interpretació del llenguatge". Potser Nietszche i Heidegger no són de l'agrat de tothom però la idea que el llenguatge determina fortament la nostra visió del món és prou coneguda. En definitiva, si la qüestió és encara oberta vol dir que es tracta de problemes difícils i fonamentals. També vol dir que els conceptes heretats (com "subjecte" i "pensament") no poden llençar-se com arguments de pes contra les posicions funcionalistes de la IA sense un exàmen crític previ.

Malgrat alguns precedents, la filosofia occidental ha tendit a qüestionar només l'estatut de l'"objecte" i no ha acabat de reconèixer la natura problemàtica del "subjecte". Una exepció que val a esmentar és Hume, que proposa que no hi ha tal cosa (no hi ha un "jo", un "subjecte"): "Quan torno la meva reflexió envers jo mateix, mai no puc percebre aquest jo sense la presència d'una o més percepcions [...]. Per tant, és llur composició la que constitueix el jo. Totes les nostres percepcions particulars són diferents i discernibles [...], i llur existència no depèn de res." [A Treatise On Human Nature, llibre 1, punt 4, sec. 7.]

Fent Zen

Tanmateix, per tal com he repetit l'adjectiu "Occidental" en parlar de la filosofia de la ment, hom pot preguntar-se si hi ha d'altres filosofies de la ment possibles. Dues escoles filosòfiques que problematitzen tant l'"objecte" com el "subjecte" són el taoisme i el budisme. En concret, el budisme zen (o ch'an en xinès) és una escola budista molt influïda pel taoisme —i força coneguda, en alguns dels seus aspectes, a Occident. Un concepte que separa el budisme de l'hinduisme és en concepte d'*anatman*. La partícula *an* és la negació i el concepte *atman* es pot traduir per "jo", "ment" o "ànima"—segons el traductor. El principi d'*anatman* és per tant el de la no-ment, el no-jo. Por semblar contradictori per una escola com el budisme que practica tècniques "mentals" com diverses formes de meditació—però la qüestió és més subtil. L'ontologia del budisme postula l'existència de *dharmes*, objectes discrets, i que la realitat es constitueix per agregats d'aquest *dharmes*. El principi d'*anatman* enuncia que la ment no és un dharma, és només un agregat. El corol·lari d'aquest principi és que *la ment no és res d'especial*, no és diferent de les altres coses del món, és un altre agregat dels components bàsics. De fet, moltes de les tècniques de "mentals" del budisme zen poden considerar-se eines per assolir el deslliurament de la il·lusió generada per aquests conceptes erronis que construeixen una imatge distorsionada del que som. Una anàlisi de la pràctica zen des del punt de vista de les neurociències és el llibre *Zen and the Brain* del neuròleg i practicant de zen James H. Austin.

La noció d'*anatman* s'assembla, només superficialment, a la proposta de l'eliminativisme. Tanmateix, el propòsit d'aquest article no és afinar distincions filosòfiques, i molt menys proposar quina escola de pensament té la raó, sigui Occidental o no. (Un llibre interessant, en aquest sentit, és *The Embodied Mind*, on Francisco Varela prova de relacionar els conceptes sobre la ment del budisme i la ciència cognitiva contemporània, inclosa la societat de la ment d'en Minsky). El propòsit d'aquest article ha estat examinar els conceptes que s'utilitzen en parlar de la IA, examinar els arguments en què habitualment s'usen aquests conceptes—i finalment intentar esbrinar perquè es tant difícil parlar i entendre's en tractar tant la IA com la "intel·ligència" en sí. Per aquest motiu l'article ha començat adreçant alguns dels arguments més habituals emprats per discutir de la "intel·ligència" en el context de la IA, per anar-s'en després a concentrar en els conceptes nuclears que s'utilitzen, i arribar finalment a les arrels filosòfiques. La conclusió general que suggereix aquest recorregut és que, lluny dels arguments habituals que menystenen les propostes de la IA, són els mateixos conceptes i arguments filosòfics sobre la ment i la intel·ligència els que són problemàtics i requereixen una reflexió més subtil i aprofundida que la realitzada fins ara.

La situació aquí i ara

En definitiva, els problemes que la IA troba avui a nivell filosòfic i de comprensió social no són diferents dels de la resta de la ciència. Els conceptes heretats sobre la ment són problemàtics tant a nivell filosòfic com de comprensió social.

A nivell de comprensió social, el rebuig a la proposició que la ment, la intel·ligència, no són més que un mecanisme és molt fort, però és essencialment igual al rebuig envers la biologia contemporània respecte del mecanisme de l'evolució i del codi genètic. Aquest rebuig envers el mecanisme (alguns en dirien "mecanicisme", però això és simplement un mot de desqualificació) es manifesta en la necessitat que "hi ha d'haver alguna cosa més". Aquesta necessitat va dur al vitalisme ara fa dos segles i als atacs revisionistes a l'evolució darwiniana. Daniel Dennet, al llibre *Darwin's Dangerous Idea*, fa una revisió als arguments que critiquen la *nova síntesi* (el paradigma que unifica l'evolució de les espècies amb l'herència genètica) sempre proposant que "hi ha d'haver alguna cosa més" en l'evolució dels sistemes vivents. La resistència a abandonar conceptes tradicionals en la conceptualització dels éssers vivents és essencialment igual a la que la IA es troba en la conceptualització de la ment i la intel·ligència.

A nivell filosòfic, els conceptes heretats en filosofia de la ment són tan "humanocèntrics" que cal llur revisió en profunditat. Deixant de banda el dualisme per raons obvies, tant el reduccionisme com l'eliminativisme no aporten gaire al debat. El reduccionisme pot tenir

una interpretació forta i una feble. La interpretació forta és que la ment es pot *reduir* al cervell i que per tant cal estudiar el cervell de manera que quan l'entenguem totalment els conceptes sobre la ment (i els conceptes de la psicologia) seran merament conceptes *derivats* del model del cervell. La interpretació feble és simplement el *fisicalisme*, on les descripcions psicològiques i de la ment tenen una correspondència amb la teoria del cervell —la diferència és que no són meres derivacions i per tant tenen una importància científica per si soles. Semblaria que la interpretació forta és poc versemblant, en primer lloc perquè l'anàloga reducció en sentit fort de la biologia a la física sembla poc versemblant, i en segon lloc perquè la reducció del software al hardware és un problema anàleg on la reducció és "possible en principi" però es fa palesa la inutilitat d'aquest reduccionisme. Conseqüentment, sembla raonable en ciència mantenir la utilitat de diversos nivells de descripció de sistemes materials, cosa compatible amb la interpretació feble del reduccionisme. Malauradament, aquest a interpretació diu poca cosa: és simplement el fisicalisme, és a dir la negació del dualisme.

Finalment, l'eliminativisme també pot tenir una interpretació forta i una feble. La interpretació forta és que *només* cal entendre el cervell i són sobrers tots els conceptes psicològics i relacionats amb la ment (conceptes "mentalístics"). Altre cop, és molt poc versemblant que el desenvolupament de la ciència vagi per aquest camí; tanmateix potser sí que la filosofia de la ment occidental hauria d'oblidar els conceptes heretats—però compte que l'eliminativisme nega que s'hagin de substituir per uns de millors, ans afirma que s'han d'eliminar i prou. La interpretació feble seria que cal oblidar els conceptes mentalístics heretats i desenvolupar nous conceptes—semblantment a com la psicologia cognitiva distingeix entre els seus conceptes i els heretats provinents de l'anomenada *folk psychology*, Ara bé, no serien mentalístics els nous conceptes? Amb quin criteri ho sabríem distingir? Sembla que l'eliminativisme ha diagnosticat els problemes existents en la filosofia de la ment però encara no és massa clar com resoldre les qüestions obertes. Daniel Dennet, un dels pocs filòsofs que s'ha pres seriosament els qüestionaments fonamentals que ha aportat la IA al debat científic, discuteix moltes d'aquestes qüestions al llibre *The Intentional Stance.*

Per últim, cal fer palès que l'objectiu d'aquest article no ha estat el "contestar" tots els arguments que s'esgrimeixen contra la IA. Per aquesta raó, no s'han tractat cap dels arguments que discuteixen si alguna cosa que no és possible en la IA d'avui ho serà o no en el futur. Els arguments de futur usualment acaben essent una qüestió empírica: La pregunta "S'arribarà algun dia a construir una IA amb intel·ligència general?" és com la pregunta "S'arribarà algun dia a construir una TOE, una Teoria del Tot?" en física. A més, de quin futur es parla? Del 2050, del 2150, o del 2550? Tampoc no s'ha tractat aquí de respondres els arguments sobre la impossibilitat, sovint basats en "experiments mentals" com l'"habitació xinesa" d'en Searle o en les idees de calculabilitat i algorísmica. Tanmateix, sí que hem discutit els conceptes que s'usen en aquesta mena d'arguments i hem vist que lluny d'aclarir els problemes són ells mateixos problemàtics. Una altra colla d'arguments argumenten contra la possibilitat de parlar d'intel·ligència en sistemes artificials fent recurs a conceptes associats, com per exemple la *consciència*. Avui per avui, el concepte de consciència sembla cada cop més usat en els arguments que busquen "alguna cosa més". En aquest article la discussió s'ha centrat sobre quan es podia (i quan no es podia) parlar d'"aspectes mentals", com intel·ligència i comprensió, en diversos sistemes materials. Per tal com la consciència és un aspecte de la ment també en podriem haver discutit, però la llargada de l'article hagués patit—especialment perquè la quantitat de malentesos i connotacions del mot "consciència" és tant gran com el de "ment". Aquesta opció no vol defugir el tema de la consciència ni vol dir que sigui un argument guanyador: les qüestions adreçades aquí sobre el concepte de *ment* són també aplicables al de *consciència*. En primer lloc, els conceptes heretats al voltant del tema de la consciència són igualment problemàtics que els conceptes al voltant del tema de la ment. En segon lloc, l'argument contra l'humanocentrisme d'aquests conceptes també s'aplica a la consciència: els primats tenen algun tipus de consciència al igual que tenen un tipus de ment (de comprensió, d'intel·ligència). Sabem potser poc encara d'aquest tema de la consciència animal, però s'en sap prou per no poder eliminar el concepte dels animals no humans. Els advocats més forts d'aquesta posició arriben a dir que la consciència és un fenomen tant diferent a tots els altres fenòmens de l'univers que no s'en pot fer (que és *impossible* fer-ne) un model

científic. El temps dirà si aquesta afirmació metafísica és correcta, de moment hi ha autors (com Baars al llibre *In the theater of consciousness*) que mostren com pot començar a fer-se un model científic d'això que sembla l'últim recurs dels advocats de "hi ha d'haver alguna cosa més".

Potser no hi ha alguna cosa més, potser la ment no és res d'especial. Hi ha un conte de Bodhidarma, l'introductor a Xina del budisme zen (ch'an), que il·lustra el concepte d'*anatman*.

> Hui-K'o, se sentia amoïnat en la seva cerca del camí (el tao). Molts cops havia suplicat Bodhidarma que l'ensenyés a assossegar la seva ment. Sempre, Bodhidarma li ho refusava, fins que finament un dia Bodhidarma li preguntà "Què cerques?". "L'assossec de la ment" responguè Hui-K'o. "Mostra'm la teva ment" digué Bodhidarma, " i jo l'assossegaré". "Però és que quan cerco la ment, no puc trobar-la" hi replicà. "Ve-t'ho aquí", diguè Bodhidarma. "És clar!" riguè Hui-K'o, i assolí el *satori*.

Bibliografia

Austin, J.H. (1998) *Zen and the Brain: Toward an Understanding of Meditation and Consciousness*. The MIT Press.
Baars, B. J. (1997) *In the theater of consciousness*. Oxford University Press.
Chalmers, D. J. (1996) *The conscious mind*. Oxford University Press.
De La Mettrie, J. O. (1748) *L'Homme Machine.*
Dennett, Daniel (1987) *The Intentional Stance.* Cambridge, Mass.: Bradford Books.
Dennet, D. (1995), *Darwin's Dangerous Idea.* Simon &Schuster.
Lakoff, G., Johnson M. (1984) *Metaphors we live by*. The University of Chicago Press.
Lakatos, I. (1976) *Proofs and Refutations*. Cambridge University Press.
McCulloch (1965) *The embodiments of mind.* MIT Press
McCulloch and W. Pitts(1943) A logical calculus of the idea immanent in neural nets. *Bulletin of Mathematical Biophysics*, p. 115-137.
Rey, G (1997) *Contemporary philosophy of mind*. Blackwell Publishers.
Simon, H. (1981) *The sciences of the artificial*. 3rd edition 1996. The MIT Press.
Varela, F. J. (1992), *The embodied mind*. The MIT Press.
Wolfram, S (2002) *A new kind of science*.Wolfram Media Inc.